Competing Against Stereotypes: Stereotyped Beliefs and Willingness to Compete*

Argun Hild[1,2] & Michael Hilweg-Waldeck[1,2]

[1]University of Mannheim
[2]ZEW Mannheim

Abstract

Career choice, earnings, and other key economic outcomes have been linked to gender differences in willingness to compete. We examine how gender stereotypes shape these differences. We conduct a meta-study of prior work and demonstrate that the wide variation in gender competition gaps can be explained by stereotypes: Men enter competitions more in traditionally male-stereotyped domains, whereas in female-stereotyped domains, the gap is smaller or even reversed. Importantly, these differences are not explained by gender gaps in performance. To explore mechanisms, we collect belief data in an elicitation experiment. We find that stereotyped beliefs about gender performance differences explain more than half of the variation in competition gaps in the literature. In follow-up experiments, we manipulate stereotypes through framing and informational cues about others' beliefs. Although these interventions significantly shift beliefs, the effects do not translate into changes in competitive behavior. Our findings highlight the importance of stereotypes in shaping gender gaps in competitiveness while suggesting that shifting beliefs alone is unlikely to close these gaps without deeper or longer-term interventions.

JEL: D91, J16, C90

Keywords: Gender, Competitiveness, Stereotypes, Beliefs, Experiment

Version: October 14, 2025

*We thank Anna Dreber Almenberg, Guido Friebel, Adrian Hillenbrand, Marion Ott, Bertil Tungodden and Alexander Cappelen for helpful comments and suggestions. We particularly thank our supervisors, Henrik Orzen and Wladislaw Mill, for their guidance and support. We also thank seminar participants at the Stockholm School of Economics, Stockholm University, and NHH Bergen, as well as participants at WEAI San Francisco, IMBESS Valencia, BEE Florence, Verona VEM, Prague PCBS, and Chicago School in Experimental Economics for their valuable feedback and discussions. Finally, we thank the lab managers Andrej Woerner and Holger Gerhardt for their assistance in running our experiments.

1 Introduction

In their seminal study, Niederle and Vesterlund (2007) found a 38 percentage point gender gap in willingness to compete: Men were significantly more likely than women to choose a competitive Tournament over a non-competitive Piece-rate compensation scheme. This result has been consistently replicated and, importantly, the stylized experimental measure of competitiveness strongly predicts key economic outcomes, including educational choices and career trajectories. Field experiments in the Netherlands (Buser et al., 2014) and Switzerland (Buser et al., 2022) have shown that girls and boys who choose competition are more likely to select competitive academic tracks. Similarly, competitiveness predicts income expectations of undergraduate students (Reuben et al., 2017). Among MBA graduates, gender differences in willingness to compete explain approximately 10% of the gender pay gap (Reuben et al., 2019). These gaps have real costs. When high-performing women opt out of competition and men with lower performance opt in, talent is misallocated. At the societal level, this distorts selection into education and career tracks, reduces efficiency, and perpetuates gender inequality.

Given the importance of competitiveness for economic outcomes, understanding whether these differences stem from innate gender differences or malleable social factors has profound implications for both theory and policy. Recent theoretical and empirical studies have proposed *gender stereotypes* as an important explanatory factor behind gender differences in economic decision making (Bordalo et al., 2016; Exley et al., 2025; Reuben et al., 2014). For instance, Dreber et al. (2014) report that girls are more willing to compete in a female-stereotyped verbal task than in a male-stereotyped math task. These results point directly to our central question: How do stereotypes shape gender differences in competitive behavior?

This paper makes three contributions. First, we provide the first systematic evidence, through a meta-study of past papers, that gender competition gaps are systematic and predictable, aligning closely with domain-specific stereotypes. This reframes scattered results in the literature as a coherent pattern, rather than inconsistent anomalies. Second, we demonstrate that actual gender performance differences explain little of this variation, while stereotyped beliefs about performance differences explain more than half of it. Third, we provide experimental evidence that stereotypes are malleable but that shifting beliefs alone does not change competitive behavior.

Crucially, prior studies measure competitiveness in domains perceived as male-stereotyped (e.g., math, stock forecasting) and female-stereotyped (e.g., verbal fluency, emotion recognition), making them ideal for studying how stereotypes influence competition. We exploit this variation via a meta-study to examine how gender differences in willingness to compete vary with domain-specific stereotypes. Drawing on data from 127 previous studies,

covering 30 distinct domains, we find a clear pattern: In male-stereotyped domains, such as forecasting stock prices or solving math problems, men compete more than women. On the other hand, in female-stereotyped domains, such as recognizing emotions from images or memorizing cards, the gap is smaller or even reversed.

An intuitive explanation for the domain-specific variation in competitiveness is that gender stereotypes might be mirroring actual gender performance differences. If men outperform women in male-stereotyped domains and vice versa, gender differences in competitiveness might naturally arise. However, we find that in most domains there are no discernible gender differences in average performance. In fact, gender performance differences explain only 1.7% of the variation in gender competition gaps.

Even if there are no gender differences in average performance, potential disparities at the top of the performance distribution could still matter. Perceptions about which gender dominates at the top are particularly relevant because tournament entry decisions depend less on average performance and more on beliefs about the performance of the strongest possible opponent. To address this, we use participant-level data from a subset of prior studies to compare the top end of the performance distribution. We find that gender performance gaps at the top are insignificant in most domains and cannot explain the observed variation in competition gaps. Together, our findings speak against actual gender performance differences as the mechanism through which stereotypes affect gender differences in competitiveness.

Although there are no gender differences in actual performance, people may still believe that men perform better in male-stereotyped domains and women in female-stereotyped domains. Therefore, we propose stereotyped beliefs—that is, exaggerated perceptions about a group's abilities (Bordalo et al., 2016)—as the main mechanism. This builds on evidence that gender stereotypes shape economic decisions through their influence on perceived ability and expected performance (Bordalo et al., 2019; Coffman, 2014). Unlike these studies, which examine the existence of such effects in specific contexts, we quantify their explanatory power for variation in competitiveness across a wide range of domains.

A useful way to situate our contribution is by analogy to overconfidence. Overconfidence has been defined as an individual's belief about own ability in excess of actual performance, and it has been shown to partially explain gender gaps in competitiveness. We argue that stereotyped beliefs operate at the group level in a structurally similar way: They are beliefs about relative performance of men and women that exceed the truth. In this sense, stereotyped beliefs can be viewed as the social analogue of overconfidence.

To test this mechanism, we conduct an online belief elicitation experiment, asking participants how they perceived the average performance of men and women across the domains identified from prior studies. While some domains can be broadly classified

as male- (e.g., math, spatial reasoning) or female-stereotyped (e.g., emotion recognition, verbal reasoning), our elicitation provides a direct and systematic measure of beliefs about gender performance differences within each domain. We find that beliefs are stereotyped: Participants believe that in male-stereotyped domains, men perform better than women and vice versa. Importantly, these beliefs strongly correlate with gender competition gaps, even after controlling for actual performance differences. Specifically, the stereotyped beliefs we elicit explain 54% ($p < 0.001$) of the observed variation in gender competition gaps across domains.

Having established the central role of stereotyped beliefs, we next ask whether these beliefs are malleable. If beliefs can be altered through subtle contextual cues, this would suggest a potential pathway for interventions aimed at reducing gender gaps. To test this, we develop an experimental task requiring participants to engage both math and memory skills, thus allowing us to frame it in either a male- or female-stereotyped manner. In a separate validation experiment, we confirm that this framing significantly alters participants' beliefs in gender performance differences, albeit with a small effect size.

Stereotypes can arise both from individuals' own beliefs about gender differences and from their perceptions of what others believe. To separately target these two channels, we employ two complementary treatment variations. First, we manipulate participants' first-order beliefs i.e., "I believe men (women) perform better," directly by framing the task as either a "Math task" or a "Memory task". Second, we manipulate participants' second-order beliefs i.e., "I believe others think men (women) perform better," to isolate the causal role of social perceptions in shaping individuals' own beliefs and thereby competitive behavior. To avoid deception, we first generated two opposing sets of average beliefs in a separate online experiment. We then used these average beliefs as the basis for our information treatment. Depending on treatment, participants were truthfully informed that a previous set of participants believed that men (women) perform better.

Our lab results show that beliefs are malleable. Under *Memory* framing, 44% of participants believe that women outperform men, while 31% think that men outperform women. Switching the framing to *Math* reduces the first proportion by 13 percentage points ($p < 0.001$) and increases the second proportion by 12 percentage points ($p < 0.001$). Similarly, informing participants that others think men (women) perform better shifts beliefs in the suggested direction. Together, these treatments offer strong evidence that gender stereotypes are context-dependent and susceptible to experimental manipulation. Furthermore, at the individual level, we replicate the same pattern that we document in our meta-study: Beliefs about male performance advantage correlate with the gender competition gap. In the subsample of participants who believe that women perform better than men, men and women are equally competitive. However, among participants who

believe that men perform better, the gender competition gap is 18 percentage points ($p < 0.001$). This is mainly driven by male competitiveness, in particular by men who believe that men perform better than women.

Treatments have no detectable effects on competitiveness, despite successfully shifting beliefs. This belief–behavior gap echoes similar findings in other fields, where changes in attitudes or intentions often fail to translate into behavior (Sheeran, 2002; Webb and Sheeran, 2006). These parallels suggest that interventions to close gender competition gaps may face similar challenges. We discuss two factors. First, our belief measures focused on average gender performance differences, whereas competitiveness could depend more on beliefs about top performers. Second, our two treatments were delivered separately, reducing their potency. Consequently, the modest effect sizes of our treatments may have been insufficient to shift behavior in the short term.

To address these points, we conduct a complementary online experiment where we elicit beliefs about gender differences both at the average level and at the top of the performance distribution. To maximize treatment effects, we combine the two previously used treatment variations. Participants in the "female-congruent" treatment encounter the task under the "Memory" framing and are informed that others believe that women perform better, while participants in the "male-congruent" treatment receive the opposite pairing. Although these enhanced treatments successfully shift beliefs (both average and top-end), we again find no effect on competitiveness. Therefore, we confirm that the belief-behavior gap observed in the lab is not due to an absence of treatment effects on beliefs about top-end performance differences.

Our study helps explain when and why gender competition gaps emerge: They align systematically with domain-specific stereotypes rather than actual ability differences. Whereas prior work has emphasized risk preferences or overconfidence as drivers of the competition gap, we demonstrate that stereotyped beliefs explain an order of magnitude more variation across tasks. This shifts the explanatory focus from individual preferences and biases to socially shared misconceptions. We show that these beliefs can be shifted through subtle interventions, though such changes do not immediately alter competitive behavior. These findings advance theories of stereotype-driven economic decisions and offer guidance for designing interventions to address gender gaps in competitiveness. Our approach is novel in combining a large-scale meta-study, targeted belief elicitation, and randomized lab and online experiments. This triangulation allows us to quantify the explanatory power of beliefs relative to alternative mechanisms and to test their causal malleability.

Our findings offer nuanced insights for policy interventions aimed at reducing gender gaps. Gender stereotypes have increasingly been recognized as critical to understand-

ing persistent gender inequalities in economic outcomes. Reflecting this, organizations are investing heavily in bias awareness initiatives, such as Harvard's Implicit Association Test requirements and New York City Police Department's implicit bias training programs (Worden et al., 2020). Our results suggest both promise and limitations for such approaches. On the one hand, we demonstrate that stereotyped beliefs can be shifted through relatively simple and subtle interventions. On the other hand, the belief–behavior gap we document suggests that interventions that achieve modest changes in beliefs, while potentially valuable, may be insufficient to induce behavioral changes. This highlights that successfully addressing gender differences in competitiveness may require interventions that go beyond changing surface-level beliefs to address deeper psychological and structural factors.

2 Meta-study

Competitiveness in the experimental literature is defined as the decision to enter a competitive Tournament rather than a non-competitive Piece-rate scheme (Niederle and Vesterlund, 2007). In the typical design, the Piece-rate option offers a fixed payment per correct answer, while the Tournament option offers a higher payment rate only if the participant outperforms the others in the group of participants they are competing with. Although it has been debated whether it reflects risk preferences, overconfidence, measurement error, or a distinct taste for competition (Gillen et al., 2019; Van Veldhuizen, 2022), this measure has been shown to possess external validity for important labor market outcomes such as career choice, income expectations, and realizations. In what follows, we adopt this standard measure and use it to systematically analyze how gender differences in competitiveness vary across the domains studied in prior experiments.

Gender differences in willingness to compete vary substantially across studies, raising the question of why some settings exhibit large gender gaps while others do not. One potential explanation is that the experimental tasks used to measure competitiveness evoke different gender stereotypes: For example, math-oriented tasks may be perceived as male-typed, while verbal or memory tasks may be perceived as female-typed. Such stereotypes could shape both performance and beliefs about performance and, in turn, the willingness to compete. The experimental literature on gender differences in competitiveness provides a unique opportunity to study this mechanism: Over 100 papers have measured competitiveness in various domains by using a wide range of tasks, from male-stereotyped tasks (e.g., throwing balls into buckets) to female-stereotyped tasks (e.g., recognizing emotions from images). We exploit this variation to examine how gender differences in competitiveness varies across stereotyped domains. Similarly, Markowsky

and Beblo (2022) conduct a meta-analysis showing that competition gaps are stronger in math than in verbal tasks. While their work focuses on estimating effect sizes, our contribution is to leverage the literature to study how competition gaps relate to actual performance differences versus beliefs about performance, which makes our approach a meta-study rather than a meta-analysis.

2.1 Data

We collect published and unpublished papers that measure competitiveness using Niederle & Vesterlund-style tournaments, focusing exclusively on studies where participants completed the task under both Piece-rate and Tournament incentives before making a competitiveness choice. Many of these papers include multiple experiments or administer the same task to different samples. For our purposes, each observation corresponds to a unique sample-task combination, i.e., a particular task administered to a particular sample. This leaves us with 155 sample-task combinations from 127 papers.

For each paper, we extract and standardize variables including task type, the number of male and female participants, male and female performance levels under both Piece-rate and Tournament-rate payment schemes. Competitiveness is measured as the proportion of male and female participants opting for the Tournament-rate scheme, and the gender competition gap is calculated as the difference between proportion of men versus women opting for the Tournament-rate scheme. Along with these variables, we collect the corresponding standard errors and p-values. For a more in-depth description of the data collection, see Appendix A.

Across all papers, we identify 30 unique tasks. For three of these, the corresponding papers did not provide data on either the performance or competition levels of men and women. For these, we conduct a replication experiment to collect the missing data (see Appendix E). Table 1 depicts a brief description of each task. These tasks range from male-stereotyped tasks such as the ball bucket task or the mental rotation task to female-stereotyped tasks such as the visual memory task or the emotion recognition task. More detailed descriptions as well as examples with solutions for each task can be found in Appendix F.

To further investigate whether gender differences in performance emerge at the top of the distribution, we obtained participant-level data for a subset of the tasks by contacting authors and accessing replication packages of the relevant papers. Due to limited availability of replication packages, we have data on the distribution of male and female Piece-rate performance only for 15 of the tasks.

Task	Description
Adding numbers	Sum five two-digit numbers
Anagram	Form words from given letter pairs
Ball bucket	Throw tennis balls into a bucket
Count numbers	Count ones in a matrix of ones and zeros
Count colors	Count white cells in a matrix of colors
Count letters	Count how many times a letter occurs in a sequence
Data search	Scrape and input data for a list of companies
Emotion recognition	Identify emotions in facial images
Find hidden words	Find words in letter matrices
Form words from letters	Form words from eight given letters
Letter difference	Identify differing letters in two matrices
Lego building	Use Lego bricks to build a specific shape
Maze	Navigate an electronic maze using the keyboard
Math-Memory	Flip cards to find hidden pairs of summations under either "Math task" or "Memory task" framing
Mental rotation	Identify the rotated version of a 3D shape
Multiplication	Multiply two given numbers
Number in numbers	Find sums in a number to match a shorter number
Quiz	Answer multiple-choice questions on various topics
Economics quiz	Answer multiple-choice questions in Economics
Rearrange words	Rearrange words to form a sentence
Rope skipping	Skip/jump the rope to earn points
Search summation	Find numbers that add up to 100 in a matrix
Sort shapes	Order six blocks with varying shapes & sizes
Spot the difference	Spot differences between two similar images
Stock forecasting	Predict stock prices based on cues and history
Typing	Type five given letters correctly
Toy fishing	Catch a prop fish with a magnet
Verify arithmetics	Verify correctness of arithmetic equations
Visual memory	Flip cards to find hidden animal images
Word in word	Form words using letters from a longer word

Table 1: Descriptions of the tasks used in the meta-study

2.2 Performance differences do not explain competition gaps

We report two main findings. First, across the tasks there is strong heterogeneity in gender competition gaps: In most tasks, men compete more than women with the gap being strongest in male-stereotyped tasks such as the stock forecasting task or the ball bucket task (Figure 1). However, in female-stereotyped tasks such as the emotion recognition task or the visual memory task, women compete more than men. Performance paints a very different picture. At the 5% significance level, 11 tasks display no discernible gender performance gap. Five tasks exhibit significant male performance advantages, while 1 task displays a significant female advantage (Appendix Figure 14).

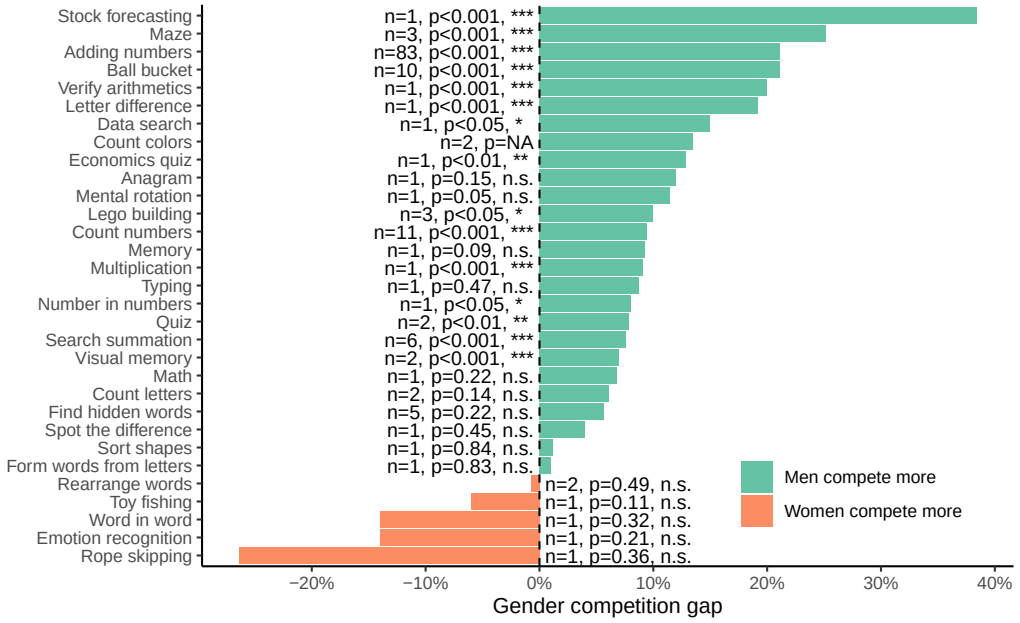


Figure 1: Gender competition gaps across tasks

The figure displays the level of gender competition gaps in the 30 tasks identified in the literature, 19 of which come from single papers. The Math-Memory task is displayed as separate tasks depending on the framing used. For each task, the gender competition gap is calculated as the difference in the proportion of men and women choosing to compete, averaged across papers. n refers to the number of papers per task. For tasks with multiple studies, p-values were combined using Fisher's method. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 2 depicts the correlation between gender competition gaps and gender differences in average performance across the 155 sample-task combinations. The correlation (corr = 0.13, p = 0.15) is small and corresponds to only 1.7% of the variation in the gender competition gaps being explained by the variation in performance differences.[1]

Beyond average performance, we also examine whether gender gaps at the top of the performance distribution correlate with competitiveness. In 5 of the 15 tasks, men are significantly overrepresented in the top decile ($p < 0.05$; Appendix Table 4), but these gaps, like average performance differences, show no meaningful correlation with competition gaps (Appendix Figure 11).

With *stereotype*, we refer to beliefs about a group that overemphasize attributes in which the group differs from a reference group, often reflecting a "kernel of truth" (Bordalo et al., 2016). It is possible that an average man has a slight, albeit undetectable, performance advantage in male-stereotyped tasks and vice versa. However, a task could be said to be stereotypical if participants hold strong and exaggerated beliefs in male (female) advantage. In the next section, we show that for most tasks, beliefs are stereotyped. We refer to beliefs that are in excess of the true gender performance differences as *stereotyped beliefs*. Since performance differences do not explain gender competition gaps, we next turn our attention to beliefs about performance differences as the mech-

[1]Unless specified, all p-values correspond to nonparametric Mann-Whitney U tests.

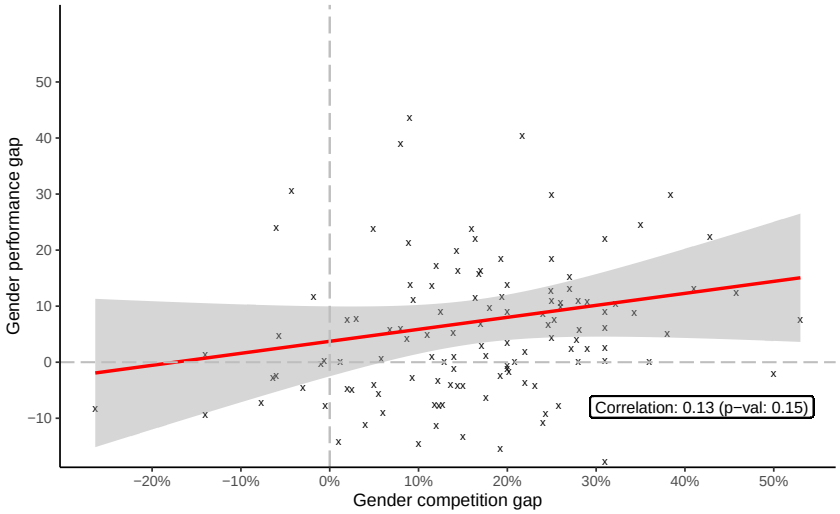


Figure 2: Scatterplot of the gender competition gaps and performance gaps

The figure depicts the correlation between gender competition gaps and gender performance gaps in published and unpublished papers in the literature. Gender competition gap refers to the difference between male and female propensity to choose the competitive Tournament-rate. Gender performance gap refers to the difference between normalized average male and average female scores in the Piece-rate stage. Each point is a single sample-task combination. The sample includes all papers including non-Western and non-adult samples. The red line shows the fitted linear regression line, and the shaded area represents the 95% confidence interval.

anism through which stereotypes explain the observed variation in gender differences in competitiveness.

3 Belief elicitation experiment

3.1 Design

Our meta-study results suggest that the gender competition gap is more pronounced in male-stereotyped tasks and that actual gender differences in average performance have little explanatory power. This raises the question whether beliefs about this difference, rather than actual differences themselves, play a central role. In particular, do participants hold stereotyped beliefs and do these beliefs help explain the observed variation in competition gaps? To answer these questions, we conduct a belief elicitation experiment measuring participants' first- and second-order beliefs about gender performance across tasks.

The experiment was conducted between 6 and 10 May 2024 with 749 participants from Prolific's U.S. sample.[2] We pre-registered two main hypotheses: (i) first-order beliefs correlate with the gender competition gaps across tasks observed in the literature, and (ii) once first-order beliefs are controlled for, second-order beliefs do not provide additional

[2]All experiments mentioned in this paper are implemented in oTree (Chen et al., 2016).

explanatory power. The analyses presented in this paper follow our pre-analysis plan.[3]

We restrict the tasks to those that were administered to Western adult samples as stereotypes and their salience may differ across cultures and age groups. We limit the number of tasks shown to each participant to a random subset of 14 out of 23, in order to reduce fatigue and improve engagement and data quality.[4] Whenever participants are evaluating a task, they see a verbal description and a picture depicting an example (see Appendix B for more details as well as a screenshot of the interface).

The experiment included attention and comprehension checks and consisted of two parts. In the first part, we elicit first-order beliefs about average gender performance levels. Here, participants are incentivized to guess how many points an average man and an average woman scored on the task. Once the participants make guesses about the 14 tasks, they move to the second part where we elicit their second-order beliefs. Here, we ask them to guess the answers of other participants to the questions in the first part. We provide no information about the gender composition of the participants whose responses they are asked to guess. One of the guesses from either part was randomly selected to determine the bonus payment. We implemented a simple linear penalty scoring rule: Participants received \$2 for perfectly accurate reports, with the bonus decreasing linearly with the absolute error of their response.[5] Participants are told that in order to maximize their bonus, they should report their guesses as accurately as possible.

3.2 Beliefs are stereotyped and predict gender competition gaps

In most tasks, there is a statistically significant belief that one gender outperforms the other. Slightly more than half of our tasks are perceived to favor men with the remaining ones perceived to favor women. As expected, participants see tasks like the ball bucket task, the stock forecasting task, or the mental rotation task as heavily favoring men. For instance, an average participant believes that men scored 25% ($p < 0.001$) more points in the ball bucket task than women, while in the emotion recognition task, they believe that women were 15% ($p < 0.001$) better than men (Figure 3). Second-order beliefs tell a similar story: Participants expect others to hold stereotypical beliefs (see Appendix B). The difference between second- and first-order beliefs reveals an interesting

[3]Pre-registration can be found at: https://osf.io/g9kmv. Secondary and tertiary questions specified in the preregistration that fall outside the scope of this text are reported in the online appendix.

[4]In total, 24 distinct tasks were used in 101 Western adult samples. For the belief elicitation experiment, we drop the form words from letters task and the quiz task on microeconomics. The Math-Memory task is treated as two separate tasks: one framed as a "Math" task and the other as a "Memory" task.

[5]Strictly speaking, this scoring rule is theoretically not incentive compatible as responses may be biased toward central or "safe" values rather than exact beliefs. However, Danz et al. (2024) emphasize that most commonly used scoring rules are behaviorally incentive incompatible, and recommend simpler rules as a more robust alternative. Following their advice, we adopted this rule, which, while theoretically imperfect, strikes a balance between clarity for participants and incentives for accurate reporting.

pattern. Across most tasks, women in particular believe that others hold stronger male-favoring beliefs than they themselves do (Appendix Figure 15).

Participants' beliefs about gender differences in performance predict the gender gaps in competitiveness observed in prior studies. Across tasks, beliefs in male advantage correlate positively with competition gaps (corr = 0.74, p $<$ 0.001) (Figure 4). Interestingly, this is driven mainly by the sensitivity of male competitiveness: Beliefs about male advantage have a significant positive correlation with male competition entry (corr = 0.52, $p < 0.01$). As expected, female competition entry displays a negative, albeit weaker and insignificant, correlation with beliefs in male performance advantage (corr = –0.25, $p \approx 0.27$) (Appendix Figure 16). This suggests that beliefs may play a more important role in increasing male competition entry than in discouraging female entry. In the subsequent lab experiment, we demonstrate that this is indeed the case.

Table 2 reports OLS regressions of believed and actual performance gaps on the observed gender competition gaps in the 23 tasks. In all columns, we control for the actual male performance advantage and interpret the belief variables as stereotyped beliefs (exceeding truth). The Performance column shows that the actual male performance advantage has a small and insignificant correlation with the gender competition gaps. More importantly, even when actual gender differences in performance are controlled for, beliefs explain a large portion of the variation in competition gaps ($R^2 \simeq 0.54, p < 0.001$). In particular, a 10 percentage point increase in perceived male performance advantage is associated with a 7.4 percentage point increase in competition gap. Overall, while actual performance differences have only a weak correlation with competition gaps, stereotyped beliefs account for 54% of the variation in gender competition gaps.

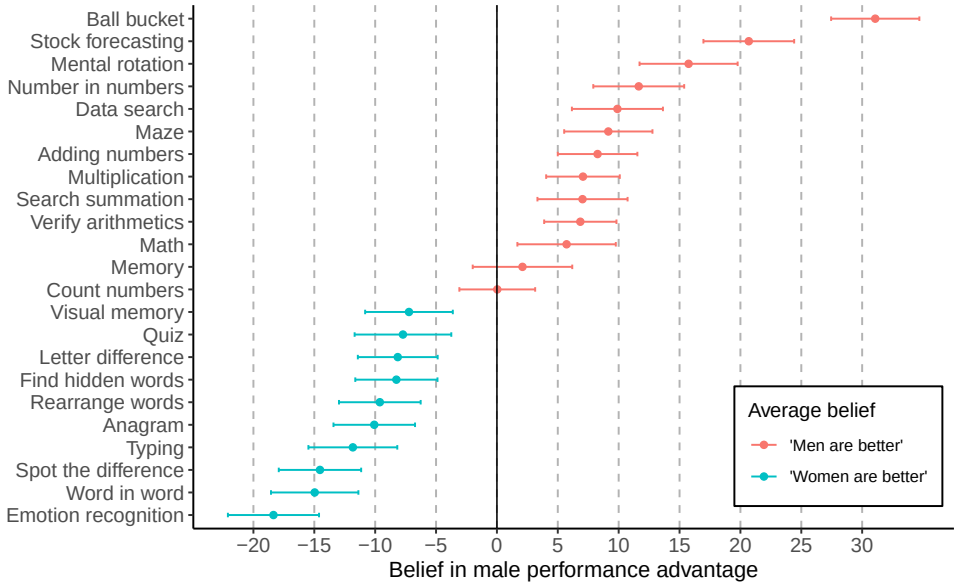


Figure 3: First-order beliefs in gender performance differences

The figure displays the level of belief in male performance advantage across tasks. The data come from the beliefs experiment. Belief in male performance advantage refers to the difference in believed average male score and believed average female score, across participants. Error bars represent 95% confidence intervals.

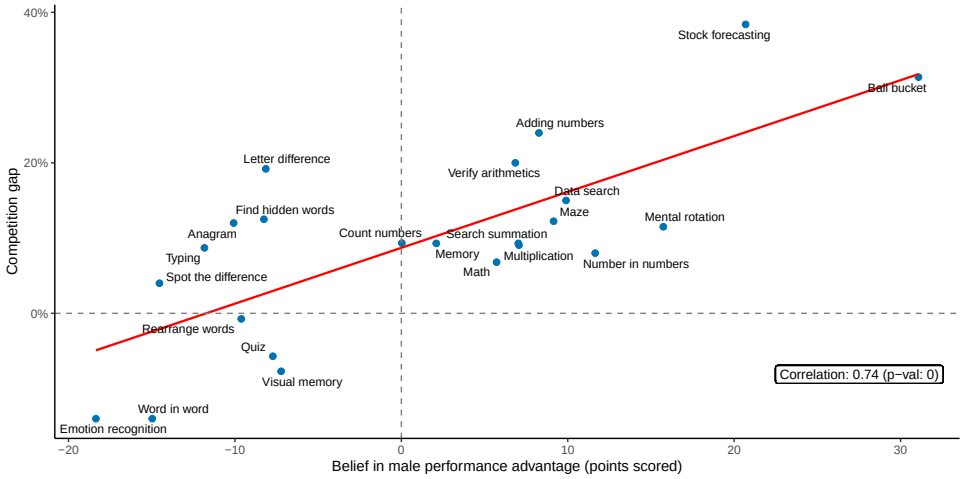


Figure 4: Scatter plot of believed gender performance and the competition gaps

The figure depicts the correlation between gender competition gaps and believed gender performance gaps across tasks. Gender competition gap refers to the difference between male and female propensity to choose the competitive Tournament-rate and comes directly from the papers utilizing the task. Belief in male performance advantage comes from our beliefs experiment and refers to the difference in believed average male score and believed average female score, averaged across participants.

Table 2: Explanatory power of beliefs

	Performance	F.O.B (1)	F.O.B (2)	S.O.B	Both beliefs
Intercept	7.668**	9.742***	9.698***	7.347**	12.270*
	(2.658)	(1.737)	(2.164)	(2.156)	(4.488)
Avg. first-order belief		0.639***	0.636**		1.326
		(0.140)	(0.171)		(1.064)
Avg. second-order belief				0.608**	−0.699
				(0.176)	(1.064)
Actual male advantage	0.353		0.006	0.087	−0.065
	(0.181)		(0.171)	(0.166)	(0.204)
R2	0.153	0.500	0.500	0.471	0.511
Num.Obs.	23	23	23	23	23

The table shows the results of a linear probability model (OLS). The dependent variable is the gender gap in each task calculated as the difference between male and female propensities to enter competition. The intercept corresponds to the residual gender competition gap. F.O.B (S.O.B) refer to first- and second-order beliefs in average male performance advantage. Avg. first- and second-order beliefs refer to mean belief in male advantage calculated as belief in average male score minus belief in average female score. Both beliefs are weighted averages of male and female beliefs. The unit of observation is the task (N = 23), and all regressors are task-level averages. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4 Lab experiment

So far, we have shown that task-related stereotypes are associated with gender differences in competitiveness: Tasks perceived to favor men exhibit larger gender competition gaps. Yet, these findings are correlational and rely on beliefs elicited from one sample to predict behavior observed in other samples. Moreover, it remains unclear whether these stereotypes are persistent or malleable. This raises three key questions: First, do participants' own beliefs about male advantage explain the competition gap within the same sample? Second, can we experimentally manipulate the stereotype associated with a task through subtle interventions? Third, does such a manipulation causally affect participants' willingness to compete? We conduct a lab experiment to address these questions.

The experiment was conducted in two university laboratories in Germany: sessions were run in Munich (November 2024) and Bonn (January 2025). In both labs, we were limited to collecting as many observations as possible within one week. We recruited 749 participants ($n_{\text{women}} = 410$, $n_{\text{men}} = 339$), in line with our pre-registered target. Due to a technical error, treatment assignment in the Munich lab was restricted to two of the four cells of the treatment matrix, while in Bonn it was balanced across all four cells (see Appendix C for more details). We follow our preregistration for all main hypotheses and analyses.[6] Some pre-registered secondary and tertiary questions fall outside the scope of this paper and are reported in the Online Appendix. In addition, we report two exploratory analyses not specified in the preregistration—the exit survey on perceived

[6]Pre-registration: https://osf.io/7duy6.

math versus memory skills, and a causal mediation analysis.

4.1 Design

4.1.1 Math-Memory task

The main objective of this experiment is to induce a variation in the stereotype associated with the experimental task. This requires a task drawing on two orthogonal skill dimensions, one regarded as favoring men, and the other one seen as favoring women. To that end, we have developed the Math-Memory task. The task starts with 16 cards (Figure 5). On the hidden side of each card is a basic summation and a colored heart (e.g., "3 + 4" and "♡"). Participants click on two cards at a time to reveal what is hidden. If the summations and the colors match, the two cards disappear and the participant earns one point.

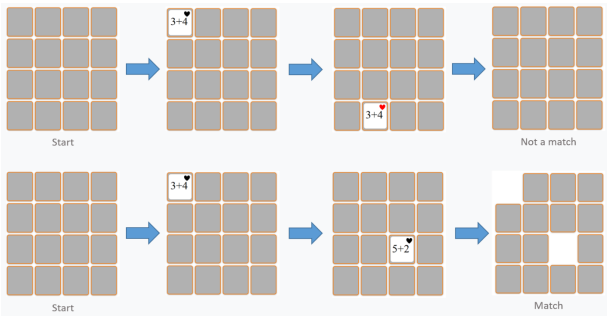


Figure 5: The Math-Memory task

4.1.2 Treatments

Stereotypes can arise either from individuals' own beliefs about gender differences or from beliefs about what others believe. Therefore, to manipulate the associated stereotype, we employ two treatment variations. The first exploits the task's reliance on math and memory skills and frames the task either as a "Math task" or a "Memory task." To strengthen our framing, we tell the participants: "Keep in mind that your performance in this task may be influenced by your math (memory) skills." If participants perceive math skills as masculine and memory skills as feminine, this stereotype priming should amplify the framing effects. In the exit survey, we elicited perceptions of math and memory skills for "a typical man" versus "a typical woman". Participants indeed rated math skills as more male-favoring and memory skills as more female-favoring.

The second treatment dimension aims at more directly influencing second-order beliefs ("what I think others think"). To vary second-order beliefs, we wanted to inform some participants that "others think men are better" while providing others with the opposite information. To avoid deception, we exploit a random artifact of our belief elicitation

experiment. In the belief elicitation experiment, we had already asked participants to evaluate the gender performance differences in the Math-Memory task under both framings. That experiment had taken 4 days. In the first two days, under both frames, the average belief was that men perform better than women whereas on the latter two days, under both frames, the average belief was that women perform better. Therefore, we selectively reveal to our participants either the first or the second piece of information. Specifically, depending on treatment, participants are shown the following information: "In a recent study on October 26 and 27 (29 and 30), we asked participants to guess the average scores of men and women in the Math (Memory) task. Result: On both days, the average participant believed that women (men) achieve higher scores than men (women)."

Most papers in the literature measure competitiveness as a binary choice to enter into the Tournament scheme versus the Piece-rate scheme. Although a binary choice has the advantage that it is a clear-cut display of participants' preferences, a continuous measure allows larger statistical power as well as more nuanced information on individual inclinations to compete. To measure competitiveness, we have adopted the method developed and validated by Saccardo et al. (2018): Participants are allocated 100 tokens which they split between the competitive Tournament-rate and the non-competitive Piece-rate scheme. Per points scored, for each token allocated to Piece-rate, the participant earns the Piece-rate and for each token allocated to the Tournament-rate, the participant earns the Tournament-rate if and only if their score is higher than the 5 other group members'. To increase comprehension of how this choice affects payoffs, we provide our participants with a scenario-calculator. This calculation displays potential earnings and admits three inputs: score, tokens invested into the Tournament, and whether the participant has the highest group score in this scenario.

4.1.3 Procedures and timeline

Appendix Figure 22 depicts the timeline of the experiment. Participants start by reading the general study instructions and complete three comprehension check questions. To ensure that participants understand they are in a mixed-gender group with five others, the instructions include an image titled "your group" which depicts three male and three female icons. We explain to the participants that the experiment contains multiple rounds and that one round will be randomly chosen to determine their bonus payment (see Appendix C for more details on incentives).

Next, we introduce participants to the task either under the "Math task" or the "Memory task" framing. Participants are told that they will perform the task under multiple rounds with each round lasting 3 minutes. After playing the task under the Piece-rate and the Tournament-rate schemes, participants move to the first Tournament

choice stage, where we elicit their competitiveness. Thereafter, they enter the first belief elicitation stage, where we elicit their first- and second-order beliefs about male and female average performance. To measure overconfidence, we ask them to guess their rank in the previous round.

After the first belief elicitation, participants receive the information intended to treat their second-order beliefs as outlined in the previous section. Thereafter, participants move to the second Tournament choice stage. Here, since we are interested in the pure effect of information, participants actively compete against the past performance of others. To measure the treatment effect on beliefs within participants, we elicit the same set of beliefs again. Since we elicit the same beliefs twice, we call them "guessing games" in both elicitation stages. In the second "guessing game", we tell participants that the reason for asking them the same set of questions again is that their beliefs may have changed due to factors such as having experienced the game in the meantime. Lastly, participants enter the third and last Tournament choice stage. There are two differences in this stage. First, after allocating their tokens between the Tournament and the Piece-rate, participants are not required to play the game again. Instead, they know that their score from the previous round will be used to determine their pay-off. Second, while in all the previous rounds they were competing against a mixed-gender group of 5 others, here we tell them explicitly that they are competing against a randomly selected participant of the opposite gender.

To elicit risk preferences, we use a version of the multiple price list (MPL) developed by Holt and Laury (2002): The safe bet and the risky alternative remain constant, but the probability in the risky alternative varies. In this MPL, participants are paid one randomly chosen choice of their 11 choices. After eliciting their risk preferences, we ask participants to evaluate the math and memory skills of "a typical man" and "a typical woman" using a 7-point Likert scale. Study instructions can be found in the Online Appendix.[7]

4.2 Results

Our main results are threefold. First, participants' beliefs about gender performance differences correlate with the gender competition gap. In fact, there is no gender competition gap in the subsample of participants who believe that women outperform men. Second, we find that stereotypes are malleable - both treatment manipulations affect beliefs about average gender performance. Third, treatments have no direct effect on competitiveness. However, mediation analysis hints at indirect causal effects through beliefs.

[7] https://github.com/ArgunHild/-Competing-Against-Stereotypes-Online-Appendix

In the whole sample, the gender competition gap is 23%. An average woman invested 43 tokens in the Tournament scheme, whereas an average man invested 53 tokens ($p < 0.01$). However, for both men and women, the distribution of tokens invested is bimodal with many choosing to go all-in on the Tournament-rate (Appendix Figure 31).

4.2.1 "Math" and "Memory" skills are stereotyped

Data from the exit survey confirm that math and memory skills are stereotyped. Participants believe that men have better math skills, whereas women have better memory skills (Figure 6). For math (memory), both male and female participants attribute a higher skill score to "a typical man" ("a typical woman") compared to its counterpart ($p < 0.001$). The proportion of participants who assign higher math skills to "a typical woman" is 10%, while 70% of participants thought that "a typical woman" has better memory skills.

4.2.2 Stereotypes are malleable

Both framing and information treatment manipulations have significant effects on beliefs. In the first belief elicitation stage, under "Math" framing, 43% of participants believed that men perform better than women and 31% believed that women perform better.[8] By contrast, under the "Memory" frame, this completely reverses: 31% think that men are better and 44% think women are better (Figure 7), a statistically significant reversal ($p < 0.001$).

Within both frames, the information "others think women are better" decreases the proportion of participants believing in male advantage ($p < 0.001$), whereas the information 'others think women are better" increases this proportion ($p < 0.001$) (Figure 8). Notably, the two treatments go hand in hand: The proportion of participants believing in male advantage is largest under the "Math" frame coupled with male-favoring information and smallest under the "Memory" frame with female-favoring information. The treatment dimensions have similar qualitative effects on second-order beliefs (Appendix Figure 27).

Across both elicitation stages, male and female participants believe that others hold stronger male-favoring beliefs than they themselves do. Similar to the meta-study findings, this effect is more pronounced among women (Appendix Figure 32). We leverage the repeated belief elicitation to examine within-participant changes. Informing participants that others think women perform better decreases second-order beliefs in male advantage ($p < 0.001$ in both frames). For first-order beliefs, however, this information decreases

[8] Treatment effects on the average participant's belief can be found in Appendix C.

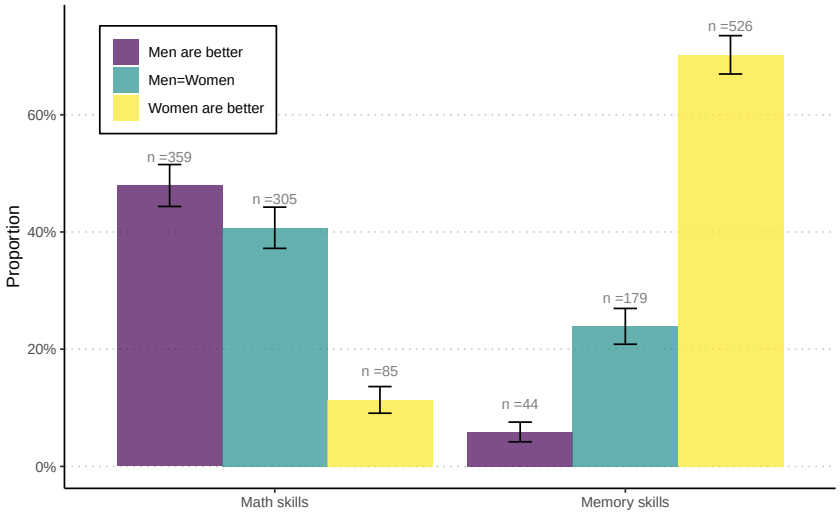


Figure 6: Perception of "Math" and "Memory" skills

Data come from an exit survey question asking participants to assess the math and memory skills of a "typical woman" and a "typical man" on a 7-point Likert scale. Error bars represent 95% confidence intervals. "Men are better" ("Women are better") belief category corresponds to those who attribute a higher score to "typical man" ("typical woman") for math (memory) skills. "Men=Women" belief category corresponds to those who have attributed the same score for the two.

beliefs in male advantage only in the Math frame ($p < 0.001$) and not in the Memory frame ($p < 0.1$) (Appendix Figure 33).

4.2.3 Gender competition gap: Only if you think men are better

In our meta-study, we show that task-level beliefs about gender performance differences explain competition gaps from previous studies. Here, we show that a similar pattern holds at the individual level. Figure 9 displays the gender competition gap across belief categories, pooling all treatments. The gender competition gap is insignificant in the subsample of participants who believe women perform better than men ($p > 0.80$). This result is consistent across treatments and all three tournament entry stages.

To our knowledge, the predictive power of beliefs in explaining the gender competition gap has so far been unexplored. Instead, the literature has focused on gender differences in risk aversion (Van Veldhuizen, 2022) and overconfidence as the two important factors. Table 3 contrasts the explanatory power of these three factors. In our sample, beliefs in male advantage exhibit a significant positive effect on competitiveness for men and an insignificant negative effect for women. This aligns with findings from our meta-study, which similarly observed that beliefs in male advantage are more strongly associated with competitiveness among men than among women.

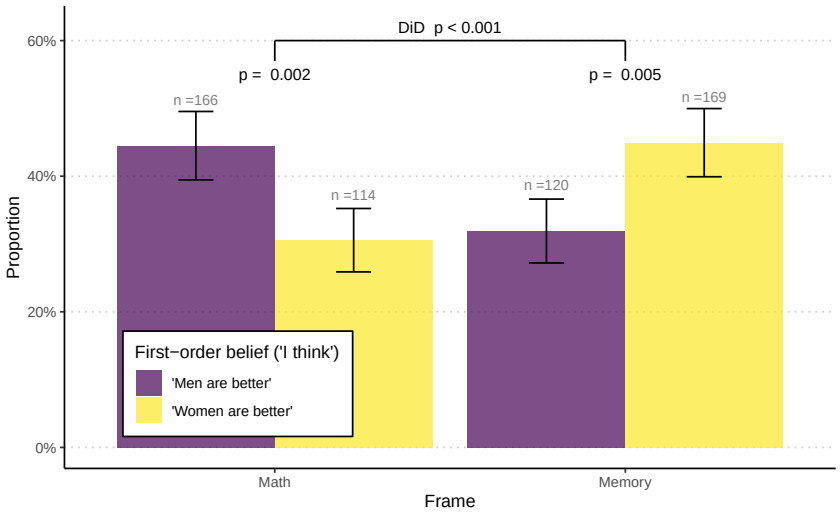


Figure 7: Framing effect on first-order beliefs (1st elicitation)

Figure displays the proportion of participants believing in male versus female advantage in average performance between "Math" and "Memory" frames. Data come from the first belief elicitation stage. Those in the "Men are better" ("Women are better") category believe that the average male (female) participant scored higher than the average female (male) participant in the Piece-rate round. The third belief category, i.e., those who believe men and women perform exactly the same is omitted for simplicity. Error bars represent 95% confidence intervals.

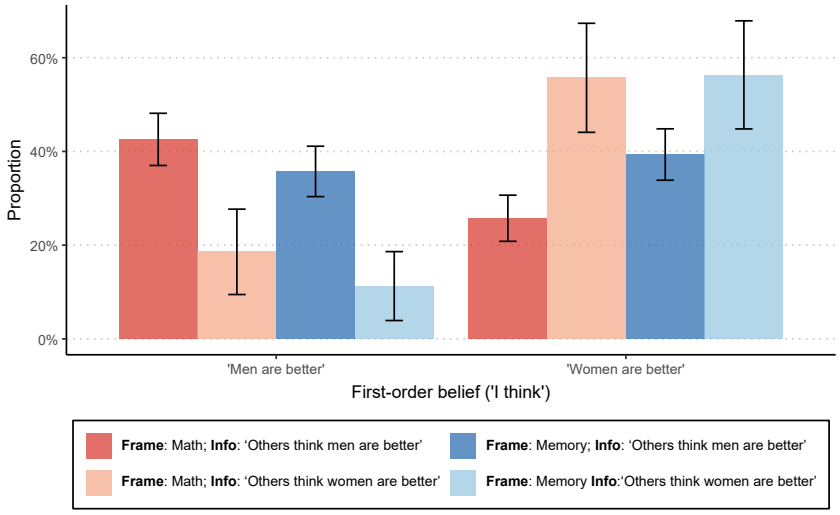


Figure 8: The effect of information treatment on first-order beliefs (2nd elicitation)

Figure displays the proportion of participants believing in male versus female advantage in average performance across all treatment conditions. Data come from the second belief elicitation stage. The proportion of participants who believe in male (female) advantage is displayed on the left (right). The third belief category, i.e., those who believe men and women perform exactly the same is omitted for simplicity. Error bars represent 95% confidence intervals.

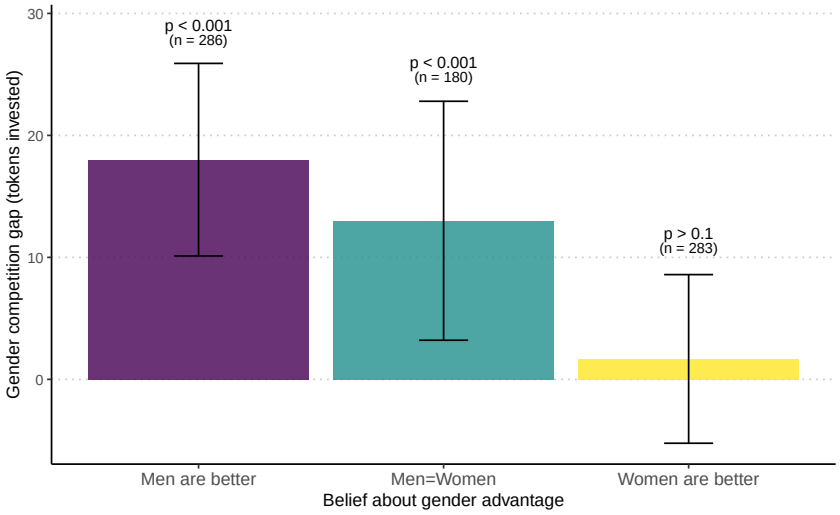


Figure 9: Gender competition gap by belief category

The figure displays the gender competition gap, measured as the difference in average tokens invested by men and women in the first Tournament choice stage. The sample is split by participants' first-order beliefs, elicited in the first belief elicitation stage. Those in the "Men are better" ("Women are better") category believe that the average male (female) participant scored higher than the average female (male) participant in the Piece-rate round.

Table 3: Determinants of competitiveness

	Base	Beliefs	Risk aversion	Overconfidence	All
Intercept	42.612***	43.142***	43.567***	41.857***	43.345***
	(1.536)	(1.556)	(1.559)	(1.606)	(1.648)
Male	10.535***	10.094***	11.046***	10.294***	10.366***
	(2.283)	(2.293)	(2.312)	(2.456)	(2.489)
F.O.B		−0.651			−0.589
		(0.346)			(0.343)
Risk aversion			−0.002**		−0.002**
			(0.001)		(0.001)
Overconfidence				1.388	1.335
				(0.873)	(0.864)
Male × F.O.B		1.561**			1.539**
		(0.586)			(0.578)
Male × Risk aversion			−0.002		−0.002
			(0.001)		(0.001)
Male × Overconfidence				−0.104	−0.050
				(1.331)	(1.312)
Num.Obs.	749	749	749	749	749
R2	0.028	0.037	0.056	0.033	0.071

The table shows the results of a linear probability model (OLS). The dependent variable is competitiveness i.e., tokens invested to the Tournament-rate in the first Tournament choice stage. F.O.B and S.O.B refer to participant's first- and second-order beliefs in male advantage respectively, calculated as belief in average male score minus belief in average female score. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.2.4 Belief-behavior gap: Treatments have no effect on competitiveness

Although both treatments influence beliefs, we do not find any direct effects on the competitiveness of either gender (see Appendix Figures 28-30). At first, it might seem puzzling that the treatment effects on beliefs do not translate into changes in competitiveness. We propose several explanations for the observed belief–behavior gap. First, there are conceptual reasons to believe that our treatments may have had offsetting effects that canceled each other out on average. Specifically, the information participants received about what others think may have caused some participants to withdraw from competition while provoking others to compete more fiercely to disprove the stereotype. In the exit survey, we asked participants to rationalize their tournament entry decisions. The responses support this interpretation: Some participants said they felt discouraged by the information that others think women perform worse, while others said that it motivated them to disprove the stereotype.

Second, although treatments change the proportion of participants believing in male advantage by about 10 percentage points, the treatment effect on the average participant's belief has a small magnitude (Appendix Figure 23, 25). For instance, under the "Memory" framing women believe that men and women attain the same score, whereas under the "Math" framing, they believe that men score 1.4 points more ($p < 0.001$). This corresponds to a modest effect size of 0.3 (Cohen's d). Therefore, although the treatments influenced explicitly stated beliefs, these shifts were likely not substantial enough to counteract a lifetime of socialization.

Third, our belief elicitation explicitly focused on perceptions of average male and average female performance. However, individuals may focus primarily on the likelihood of outperforming the strongest of the competitors rather than considering the average performance across all competitors. Therefore, although our treatments shifted beliefs about average performance, they might not have influenced beliefs regarding women's likelihood of being top performers, which could be the more relevant belief driving competitive behavior. Last, the lab experiment's measure of competitiveness differs from the canonical binary choice used in the literature. Although our measure increases power due to its continuous nature, it is possible that its complexity introduces a measurement error that masks treatment effects.

Nevertheless, because treatments significantly shifted beliefs that are themselves associated with competitiveness, treatments could have influenced competitiveness indirectly via beliefs. Mediation analysis supports this interpretation: Across all three tournament choice stages, both treatment dimensions exhibit average causal mediation effects (ACME) on competitiveness, robust to controls for risk aversion, overconfidence, and performance. However, these indirect effects are small. For example, the ACME for the framing effect

is estimated at 0.9 ($p \approx 0.03$), corresponding to a standardized effect size of 0.03 (Cohen's d), which is considered very small. It is also important to acknowledge that mediation analysis cannot rule out the possibility of omitted variables that may confound the relationship between the mediator and the outcome. A detailed discussion of the validity of the ACME estimates and full mediation results are provided in Appendix C.3. Taken together, these analyses suggest that while our treatments may have had very small indirect effects via beliefs, they were not strong enough to alter competitive behavior.

5 Online experiment

To address concerns that (i) average beliefs may not be the relevant margin, (ii) treatment effects in the lab were modest and staggered, and (iii) our competitiveness measure differed from the canonical binary choice, we conduct a complementary online experiment. The experiment is a simplified version of the lab experiment aimed at an online participant pool and differs in three key ways. First, we include a belief elicitation stage about gender differences at the top. Participants are informed that the experiment includes 300 men and 300 women and are incentivized to guess the number of men versus women among the top 50 performers. To maximize treatment effects, we merge the two treatment variations: Participants in the "female-congruent" treatment are exposed to the task under the "Memory" framing and are informed truthfully that previous study participants thought that women perform better than men. Those in the "male-congruent" treatment are exposed to the opposite configuration. Lastly, we include a binary competitiveness measure where participants choose between applying either the noncompetitive Piece-rate or the competitive Tournament-rate to their first-round score.

The experiment was conducted on 17 July 2025 with 300 female and 300 male participants from Prolific's U.S. sample. We fully follow our pre-registered analysis plan.[9] The experiment has three main results. First, similar to the lab experiment results, treatments significantly affect beliefs about average male performance advantage (Appendix Figure 37) and beliefs about how many women are among the top 50 performers (Figure 10). For instance, in the "male-congruent" treatment, the average female participant thinks that 49% of the top performers are women. The "female-congruent" treatment increases this belief to 57% ($p < 0.001$). Furthermore, as expected, merging the two treatments leads to larger effect sizes on beliefs with an average Cohen's d of 0.68.

However, despite larger shifts in beliefs, treatments again show no effects on competitiveness. In contrast to the 23% competition gap we observe in the lab experiment, here, we find no gap. Men and women invest similar amounts into the competitive Tournament-

[9]Pre-registration can be found at https://osf.io/g9kmv.

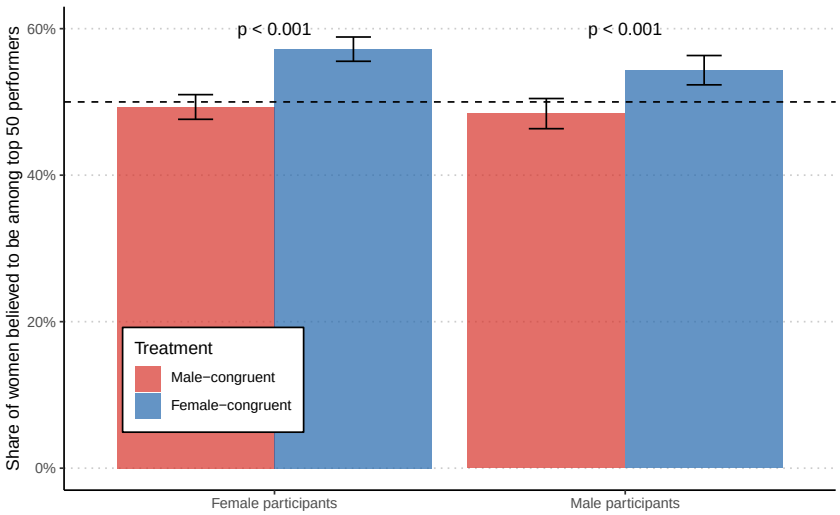


Figure 10: Treatment effect on beliefs (Online experiment)

Figure displays the treatment effects on the share of women believed to be among the top 50 performers, separated by gender of the responder. The dashed line represents an equal split i.e., the belief that there are 25 men and 25 women among the top 50 performers. Error bars represent 95% confidence intervals.

rate (≈42 tokens) and choose the Tournament-rate at similar rates (≈36%). This aligns with meta-analytic evidence that online experiments tend to produce smaller competition gaps than lab experiments (Markowsky and Beblo, 2022). Unlike in the lab experiment, here we do not find evidence that beliefs correlate with competitiveness (Appendix Table 10). However, because there is no gender competition gap at the group level, there is little scope for beliefs to explain gender differences in competitiveness.

Lastly, unlike in the lab experiment, here we do not find evidence that beliefs correlate with competitiveness (Appendix Table 10). We interpret this null result cautiously; In the absence of a gender competition gap in the online setting, there is little variation left for beliefs to explain. This does not undermine our main finding that stereotyped beliefs matter, but instead highlights how the strength of the beliefs–competitiveness link depends on the presence of a competition gap and the setting.

6 Conclusion

Gender differences in competitiveness have been linked to socially significant phenomena, including women's underrepresentation in male-dominated fields such as STEM, the shortage of women in leadership roles, and the gender pay gap. Yet, the gap is not universal: There is a large heterogeneity in the magnitude and direction of competition gaps reported in the literature. A central debate in the literature is whether the competition gap reflects a true gender difference in taste for competition or whether it can be explained by fundamental gender-specific characteristics such as risk aversion and overconfidence.

In our meta-study, we show that these gaps are not random but track domain-specific

stereotypes in predictable ways. This reinterpretation unifies two decades of evidence under a single explanatory mechanism. Specifically, in male-stereotyped domains (e.g., solving math tasks), men are far more competitive than women; in female-stereotyped domains (e.g., recognizing emotions from pictures), the gap shrinks or reverses. We show that actual gender differences in performance cannot account for this heterogeneity.

Instead, we identify *stereotyped beliefs* as the key mechanism. Participants in our belief elicitation experiment systematically overestimate gender performance differences in ways that align with task stereotypes, believing men excel in male-typed domains and women in female-typed ones. These exaggerated beliefs explain 54% of the competition gaps across studies.

We complement these findings with a lab experiment. The beliefs of our participants about average gender performance differences predict the gender competition gap, confirming that the patterns observed across studies also hold at the participant level. Previous literature has focused on overconfidence and risk aversion to explain the gender competition gap. The results of our meta-study and lab experiment suggest that beliefs about gender performance differences should be considered a primary explanatory variable as well. In this sense, stereotyped beliefs extend the overconfidence logic from the individual to the group level, but with far greater explanatory power.

We show that simple interventions, such as task framing and informational cues about what others believe, are sufficient to substantially shift participants' beliefs about gender-specific performance. We document a belief-behavior gap: Despite strong effects on beliefs, interventions aimed at changing beliefs do not lead to changes in competitive behavior. This gap mirrors findings in psychology and suggests that surface-level belief change is insufficient to alter entrenched behavioral patterns such as stereotypes. It is intuitive to think that if people no longer held biased beliefs, gender disparities in outcomes would diminish. Reflecting this, numerous educational programs, corporate training, and policy initiatives aim to reduce gender gaps by targeting stereotyped beliefs. Our results suggest that belief-based interventions can effectively shift perceptions, but we find no evidence of immediate behavioral change in our experimental setting. However, belief shifts might matter only over time and across cohorts as persistent exposure to stereotype-challenging frames could change norms even if one-off interventions fail.

Our findings resolve one puzzle but open another: While stereotypes align closely with gender differences in competitiveness, are they truly a cause? One interpretation is that stereotypes originate in historical accidents and institutional legacies, and then shape both beliefs and preferences simultaneously. If so, interventions targeting explicit beliefs alone may be insufficient, because stereotypes operate through deeper cultural channels, including identity and belonging. Another interpretation is that stereotypes reflect un-

derlying differences in preferences for domains rather than competitiveness per se. In this view, the so-called general "taste for competition" may instead be domain-specific: Individuals compete where they feel aligned with the stereotype, and withdraw where they do not. Just as overconfidence explains individual decisions, stereotyped beliefs may be the group-level analogue shaping domain-specific competitiveness. This reframes competitiveness not as a stable trait but as a context-dependent behavior shaped by stereotypes. Distinguishing between these interpretations remains an open challenge, and addressing it is essential for a full understanding of how gender differences in competitiveness arise. Taken together, our evidence leans strongly toward nurture-based explanations, namely, that gender differences in competitiveness arise less from innate differences and more from socially shared misperceptions. This suggests that what often appears as a fundamental gender gap may, instead, be a product of stereotypes that shape behavior across contexts.

References

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51*(6), 1173.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–73.

Brenøe, A. A., L. Heursen, E. Ranehill, and R. A. Weber (2022). Continuous gender identity and economics. In *AEA Papers and Proceedings*, Volume 112, pp. 573–577. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Buser, T., M. Niederle, and H. Oosterbeek (2024). Can competitiveness predict education and labor market outcomes? evidence from incentivized choice and survey measures. *Review of Economics and Statistics*, 1–45.

Buser, T., N. Peter, and S. C. Wolter (2022). Willingness to compete, gender and career choices along the whole ability distribution. *Experimental Economics 25*(5), 1299–1326.

Chen, D. L., M. Schonger, and C. Wickens (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics 129*(4), 1625–1660.

Danz, D., L. Vesterlund, and A. J. Wilson (2024). Evaluating behavioral incentive compatibility: Insights from experiments. *Journal of Economic Perspectives 38*(4), 131–154.

Dreber, A., E. Von Essen, and E. Ranehill (2014). Gender and competition in adolescence: task matters. *Experimental Economics 17*, 154–172.

Exley, C. L., O. P. Hauser, M. Moore, and J.-H. Pezzuto (2025). Believed gender differences in social preferences. *The Quarterly Journal of Economics 140*(1), 403–458.

Gillen, B., E. Snowberg, and L. Yariv (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy 127*(4), 1826–1863.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309.

Kray, L. J., L. Thompson, and A. Galinsky (2001). Battle of the sexes: gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology 80*(6), 942.

MacKinnon, D. P., A. J. Fairchild, and M. S. Fritz (2007). Mediation analysis. *Annual Review of Psychology 58*(1), 593–614.

Markowsky, E. and M. Beblo (2022). When do we observe a gender gap in competition entry? a meta-analysis of the experimental literature. *Journal of Economic Behavior & Organization 198*, 139–163.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Reuben, E., P. Sapienza, and L. Zingales (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences 111*(12), 4403–4408.

Reuben, E., P. Sapienza, L. Zingales, et al. (2019). Taste for competition and the gender gap among young business professionals. *NYU Abu-Dhabi Working Paper 31*.

Reuben, E., M. Wiswall, and B. Zafar (2017). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal 127*(604), 2153–2186.

Saccardo, S., A. Pietrasz, and U. Gneezy (2018). On the size of the gender difference in competitiveness. *Management Science 64*(4), 1541–1554.

Sheeran, P. (2002). Intention—behavior relations: a conceptual and empirical review. *European Review of Social Psychology 12*(1), 1–36.

Shrout, P. E. and N. Bolger (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods 7*(4), 422.

Tingley, D., T. Yamamoto, K. Hirose, L. Keele, and K. Imai (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software 59*, 1–38.

Van Veldhuizen, R. (2022). Gender differences in tournament choices: Risk preferences, overconfidence, or competitiveness? *Journal of the European Economic Association 20*(4), 1595–1618.

Webb, T. L. and P. Sheeran (2006). Does changing behavioral intentions engender behavior change? a meta-analysis of the experimental evidence. *Psychological Bulletin 132*(2), 249.

Worden, R. E., S. J. McLean, R. S. Engel, H. Cochran, N. Corsaro, D. Reynolds, C. J. Najdowski, and G. T. Isaza (2020). The impacts of implicit bias awareness training in the nypd. *The John F. Finn Institute*.

Zhao, X., J. G. Lynch Jr, and Q. Chen (2010). Reconsidering baron and kenny: Myths and truths about mediation analysis. *Journal of Consumer Research 37*(2), 197–206.

A Meta-study

A.1 Data collection

Following the methodology of Markowsky and Beblo (2022), we collect all working and published papers by systematically searching for recent research on gender differences in competitiveness. On November 20, 2023, we queried Google Scholar and Web of Science using the search terms "compet* AND gender" and "tournament AND gender." Additionally, we retrieved the papers citing Niederle & Vesterlund (2007) using Google Scholar. This process yielded around 20.000 papers. We kept only the papers that measured competitiveness using a Niederle & Vesterlund style tournament.[10] Some papers administer the same task to multiple samples, e.g., adults and teenagers, whereas some other papers administer different tasks to the same sample. Since we are interested in the performance and competitiveness of past study participants in particular tasks, a sample-task combination constitutes an observation of which we have 155 in general. Restricting the data to Western adult samples which yields 101 observations from 88 papers.

All variables mentioned in the paper have been manually scraped. For each observation, the main variables pertaining to the tasks are: sample size, number of women and number of men in the sample, task type (i.e., task name) and task description as scraped from the paper.

For each observation, our main variables for performance measures are raw performance (i.e., number of tasks correctly solved) of both men and women in both Piece-rate and Tournament-rate settings wherever possible as well as their standard errors and associated p-values. Where possible and necessary, we computed the p-values from the standard errors and vice versa. In these calculations, if the number of male versus female participants was not specified, we assumed that the sample was gender-balanced i.e., $n_{male} = n_{female}$. We normalized the average performances such that the average performance of men and women added up to 200.[11] In other words, in each task, the average participant score was normalized to 100.

Our main variables of competitiveness are the proportion of men and women choosing competition, the raw competition gap as the difference in gender-competition-propensities, its standard error, and the associated p-value. Whenever the standard error or the p-value was not reported, we manually computed it from the available data. Similarly, if the exact number of male versus female participants was not specified, we assumed that the sample

[10] Some papers measure competitiveness using survey questions or via Piece-rate equivalence or through multiple price lists. We exclude these papers to maintain comparability across papers.

[11] For example, if in a task average male score was 12 and average female score was 15, then the scores were normalized to $12 * \frac{200}{12+15}, 15 * \frac{200}{12+15}$, respectively. This normalization preserves the percentage difference between the average scores of both genders.

was gender-balanced. Some papers report only the residual gender competition gap that is left after regressing competitiveness on the gender dummy and other covariates and not the raw competition gap. In these cases, if the proportion of men versus women self-selecting into competition is available, we manually calculated the raw competition gap. The results reported in the paper only utilize the raw competition gap, since the residual competition gap depends very much on the covariates included in the regression and the papers vary in which covariates they include.

To assess the statistical significance of gender differences in competition propensity for each task, we adopted a hierarchical approach to ensure that p-values are consistently available. First, we used the p-values reported in the original studies. For studies lacking reported p-values but providing the mean difference in competition propensity and the corresponding standard error, we computed approximate p-values under the assumption of normality. If neither source was available, but information on the proportion of men and women choosing to compete along with sample sizes existed, we computed p-values using a two-proportion z-test. Finally, in cases where none of the information above was available, we used residual p-values reported in the original studies. When multiple papers contributed data for the same task, we combined individual p-values using Fisher's combined probability test, which aggregates evidence across studies while accounting for the contribution of each study's significance. For tasks where neither p-values nor sufficient information to compute them manually were available, no p-value is reported.

Fisher's combined probability test is a meta-analytic method that allows the synthesis of independent p-values from multiple studies that examine the same hypothesis. Specifically, the test statistic is calculated as the sum of the natural logarithms of each individual p-value, multiplied by -2. Under the null hypothesis of no effect, this test statistic follows a chi-squared distribution with degrees of freedom equal to twice the number of combined studies. This approach efficiently integrates the statistical evidence from different sources, giving more weight to studies with highly significant results, while accommodating the variability across studies.

A.2 Gender performance gap at the top

For each task, we rely on a single corresponding paper from which we retrieved data to measure gender performance gaps at the top decile. For each task, we searched for the replication packages of corresponding papers and obtained individual-level performance data for 14 of the 22 tasks.[12] Table 4 reports performance gaps at the top decile. Most tasks do not exhibit significant gaps at the top. Count numbers and memory tasks exhibit

[12]In addition, we contacted authors for data. We thank Keana Richards, Enzo Brox, and Olga Shurchkov for sharing their data.

a male performance advantage significant at $p < 0.001$ and the search summation task displays a male performance advantage significant at $p < 0.01$.

A.3 Other figures and tables

Table 4: Gender performance gaps at the top 10%

Task	Men (%)	Women (%)	Gap (pp)	p-value	
Adding numbers	62.9	37.1	25.7	0.175	
Count numbers	65.9	34.1	31.8	0.004	**
Data search	63.8	36.2	27.7	0.079	
Search summation	64.6	35.4	29.2	0.025	*
Visual memory	36.8	63.2	-26.3	0.359	
Quiz	45.0	55.0	-10.0	0.824	
Spot the difference	55.6	44.4	11.1	1.000	
Math	50.0	50.0	0.0	1.000	
Memory	77.1	22.9	54.3	0.002	**
Number in numbers	75.0	25.0	50.0	0.289	
Word in word	69.2	30.8	38.5	0.267	
Mental rotation	66.7	33.3	33.3	0.122	
Multiplication	67.6	32.4	35.3	0.000	***
Find hidden words	25.0	75.0	-50.0	0.146	
Maze	82.4	17.6	64.7	0.013	*

This table reports the gender composition of the top 10% of performers in each task, defined as participants scoring above the 90th percentile of the pooled performance (Piece-rate) distribution, separated by gender. The "Men (%)" and "Women (%)" columns show the proportion of top performers who are men or women, respectively. The "Gap (pp)" column displays the difference in these shares. The p-values come from binomial tests of proportions under the null hypothesis of equal representation. For tasks without a forced Piece-rate stage (Find Hidden Words and Multiplication), performance data are pooled across both Tournament and Piece-rate participants. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

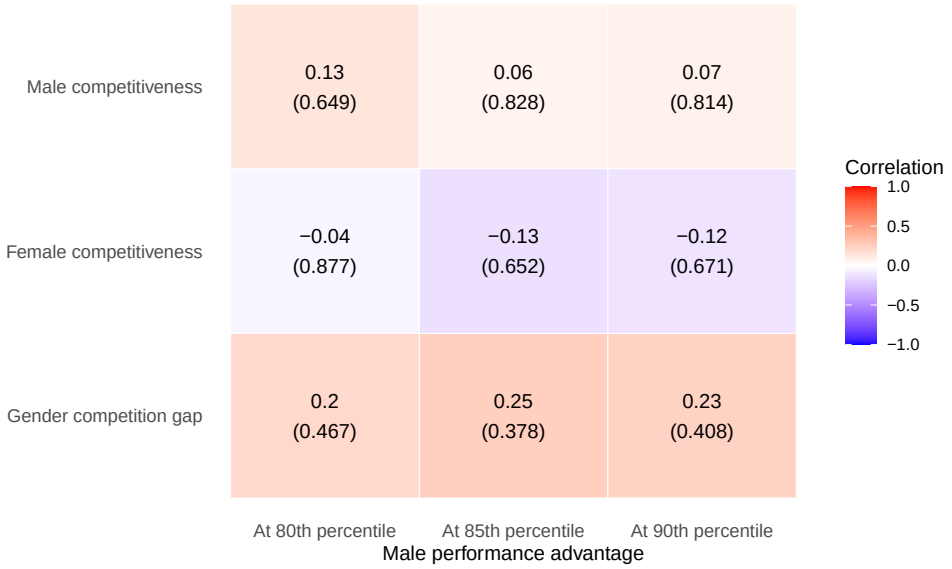


Figure 11: Correlation heatmap: competitiveness and gender performance gap at the top

The figure displays the correlations between gender performance gaps at different percentiles (80th, 85th, 90th) and three measures of competitiveness: male competitiveness, female competitiveness, and the gender competition gap. The first, second, and third columns, correspond to male performance advantage at the 80th, 85th, and 90th percentile, respectively. Male performance advantage at each percentile is calculated as the difference between share of men versus women above the Piece-rate performance respective cutoff. Each cell reports the Pearson correlation coefficient and the associated p-value (in parentheses). The correlations are calculated across tasks, and no correction is applied for multiple testing. The performance data for each task comes from a single paper while competition data is averaged across all available papers.

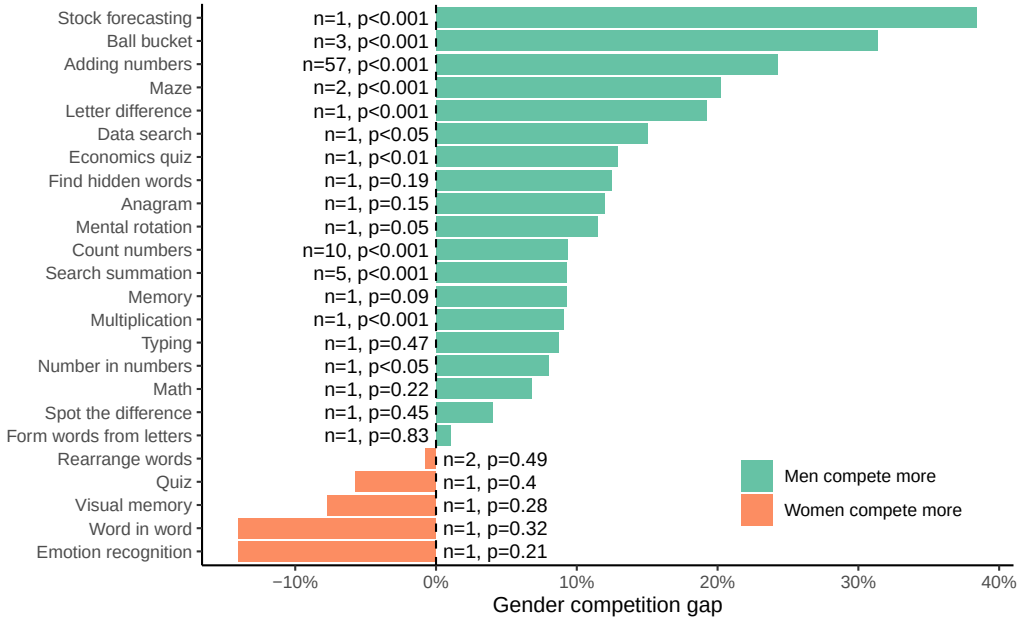


Figure 12: Gender competition gaps across tasks (Western adult samples)

The figure displays the level of gender competition gaps in the 24 tasks identified in papers utilizing Western adult samples. The Math-Memory task is displayed as separate tasks depending on the framing used. For each task, the gender competition gap is calculated as the difference in the proportion of men and women choosing to compete, averaged across papers. n refers to the number of papers per task. For tasks with multiple studies, p-values were combined using Fisher's method. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

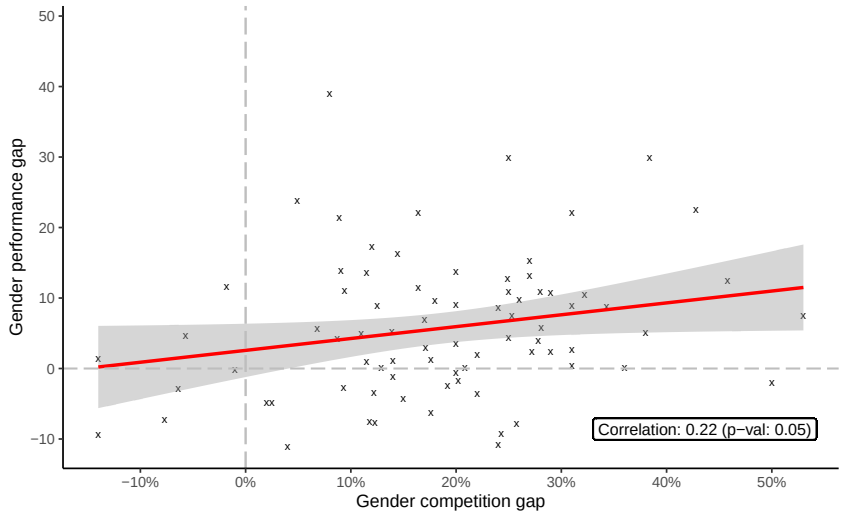


Figure 13: Scatterplot of the gender competition gaps and performance gaps (Western adult samples)

The figure depicts the correlation between gender competition gaps and gender performance gaps in published and unpublished papers in the literature. The sample of papers is restricted to those using Western adult samples. Gender competition gap refers to the difference between male and female propensity to choose the competitive Tournament-rate. Gender performance gap refers to the difference between normalized average male and average female scores in the Piece-rate stage. Each point is a single sample-task combination. The sample includes all papers including non-Western and non-adult samples. The red line shows the fitted linear regression line, and the shaded area represents the 95% confidence interval.

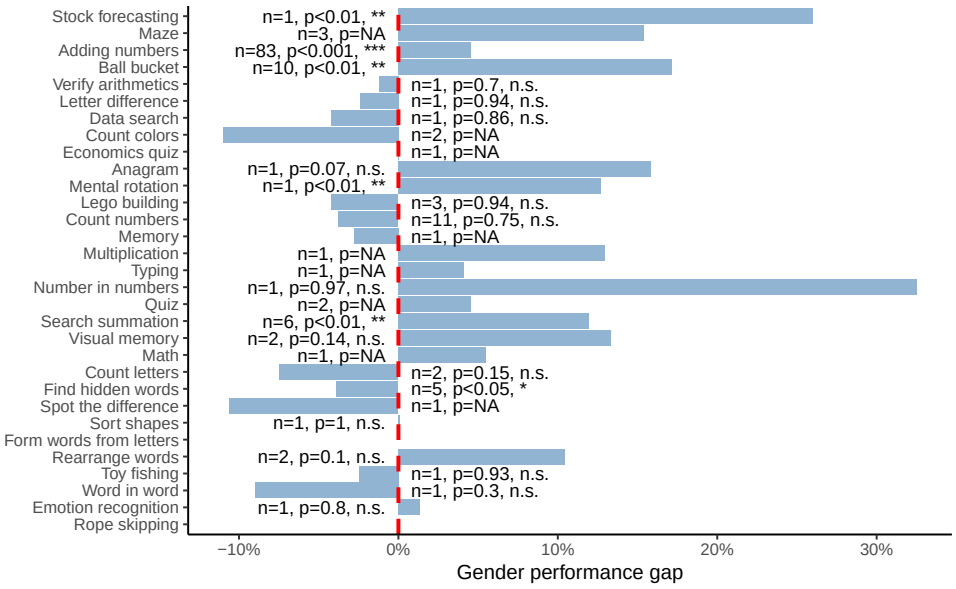


Figure 14: Gender performance gaps across tasks

The figure displays the level of gender performance gaps in the 30 tasks identified in the literature. The Math-Memory task is displayed as separate tasks depending on the framing used. For each task, the gender performance gap is calculated as the difference in the normalized score attained by men and women in the Piece-rate stage, averaged across papers. n refers to the number of papers per task. For tasks with multiple studies, p-values were combined using Fisher's method. For 8 tasks, the corresponding papers are missing both p-values and standard errors. For 3 tasks, data on performance altogether is missing.

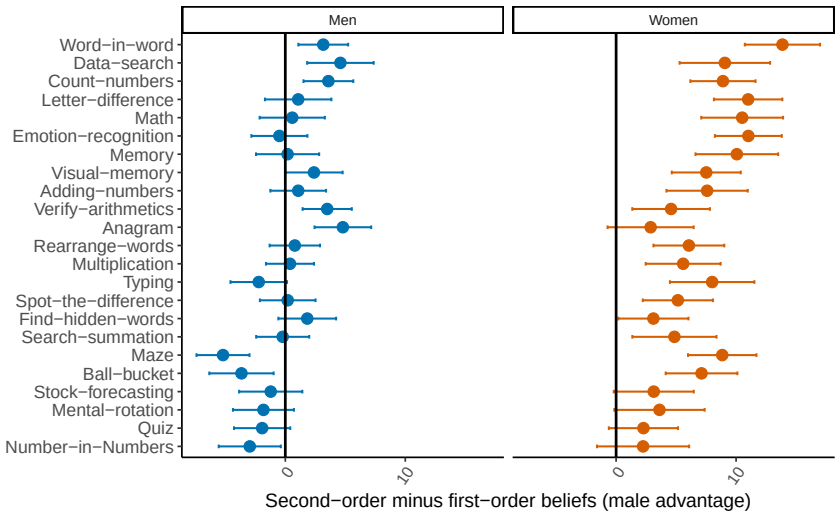


Figure 15: Difference between second- and first-order beliefs

Figure displays the average difference between second- and first-order beliefs across tasks, calculated as the participant's second-order belief in male advantage - participant's first-order belief in male advantage in that task. The left (right) panel displays the difference in beliefs for male (female) participants. Error bars represent 95% confidence intervals.

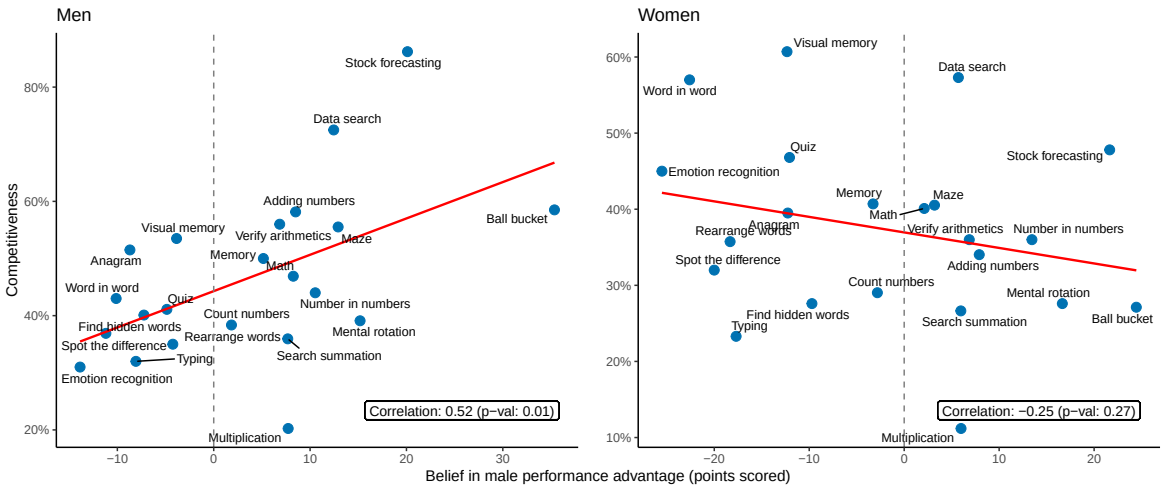


Figure 16: Scatter plot of beliefs and competitiveness by gender

The figure depicts the correlation between competitiveness and believed gender performance gaps across tasks separated by gender. Competitiveness refers to the the proportion of men (women) choosing the competitive Tournament-rate over the noncompetitive Piece-rate comes directly from the papers utilizing the task. Belief in male performance advantage comes from our beliefs experiment and refers to the difference in believed average male score and believed average female score, averaged across participants.

B Beliefs experiment

B.1 Timeline

The timeline is depicted in Figure 17. First, participants are explained that they will see 14 tasks in which past study participants have taken part. We explain to them that the study is comprised of two parts. In the first part, they see 14 randomly selected tasks out of 22 in a randomized order. On these pages, they see the title and description of the task as well as an example problem. We elicit their first-order beliefs by asking them to guess how many points an average man versus an average woman earned (Figure 18). Once participants have gone through the 14 tasks, they land on the instructions page for the second part. Here, we explain to them that their job now will be to guess the answers of other Prolific participants in the first part. In this part, they see the same 14 tasks in the same order and report their second-order beliefs similarly. We use a competitiveness measure developed by Buser et al. (2024), - a survey question reading "How competitive do you consider yourself to be." Participants answer this question by choosing a number on a 10-point Likert scale. We also adopt a continuous gender measure developed by Brenøe et al. (2022)- a survey question asking them to place themselves on a 10-point Likert scale from "very masculine" to "very feminine." The experiment included three comprehension check questions and 3 attention check questions.

B.2 Figures

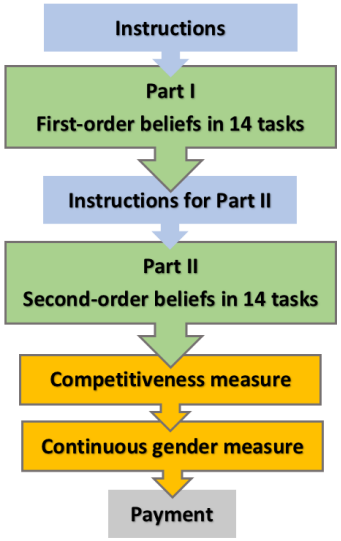


Figure 17: Timeline of the belief elicitation experiment

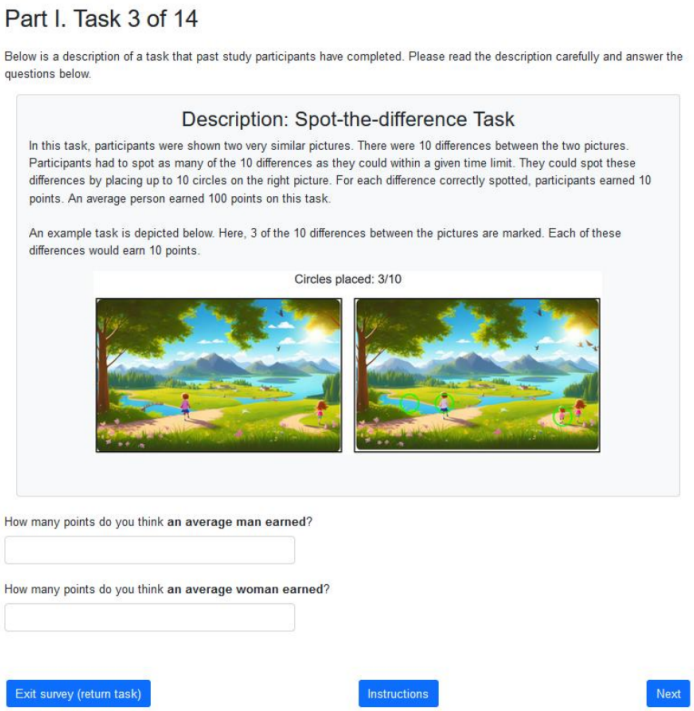


Figure 18: Elicitation of first-order beliefs

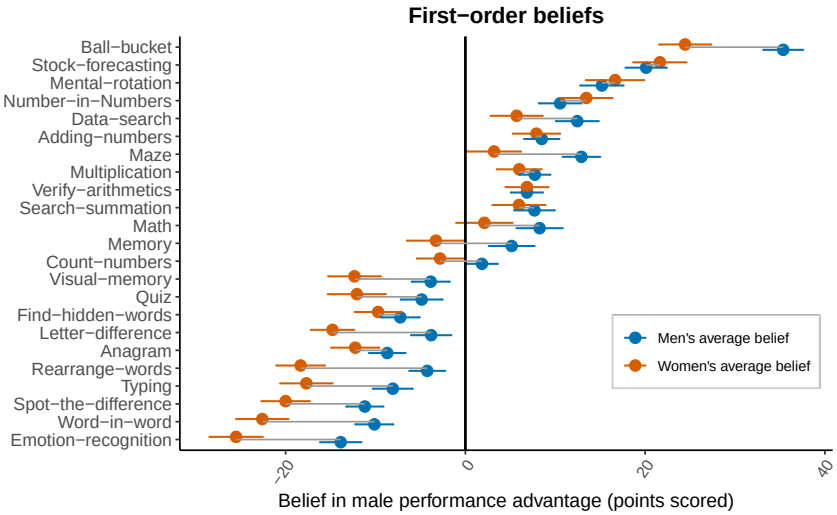


Figure 19: First-order beliefs in male advantage across tasks (men and women)

Error bars represent 95% confidence intervals.

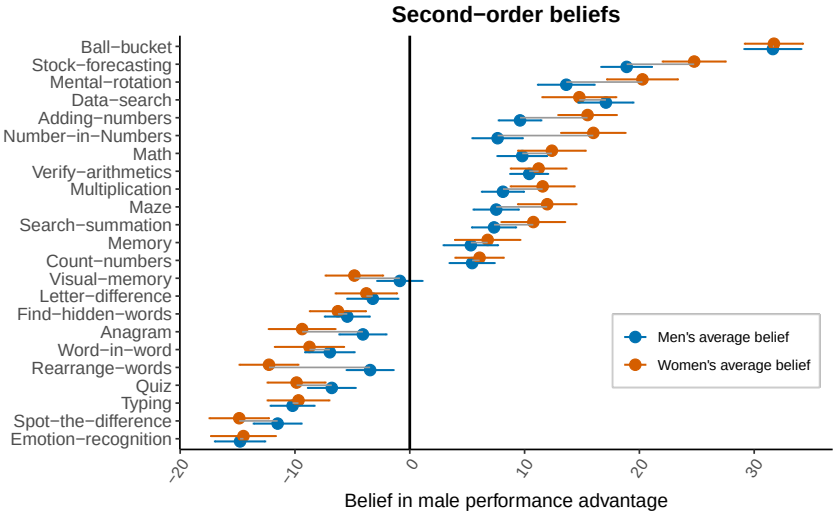


Figure 20: Second-order beliefs in male advantage across tasks (men and women)

Error bars represent 95% confidence intervals.

C Lab experiment

We were limited to collecting as many observations as possible within one week each per lab. Our total sample (n=749; 287 participants from Bonn; 462 participants from Munich) is within the indicated target range of our preregistration. The experiment consists of two treatments with two arms, resulting in a 2x2 treatment matrix (Figure 21). However, due to a coding error, in the Munich lab, the participants only received the information that others think men are better. Therefore, in our Munich sample, participants were randomly assigned to the first cell ("Math frame" and "others think men are better") or the third cell ("Memory frame" and "others think men are better") of the treatment matrix, and in the Bonn sample, the assignment was random between the 4 treatments. However, ex-post power calculations suggest no changes in the significance levels of our results in case of no coding error given our observed effect sizes.

The experiment included 3 comprehension check questions. If a participant failed a comprehension question twice, the computer alerted the experimenter so that the question can be clarified. In total, this happened for 68 participants.

Participants were told that one of the seven rounds will be randomly chosen to determine their payment. These seven rounds include the Piece-rate and Tournament rounds, the three competition choice rounds and the two belief elicitation rounds. If one of the belief elicitation rounds were chosen, participants earned a 10 EUR bonus if the answer is within 10% of the correct value. If the overconfidence elicitation was chosen, participants earned the 10 EUR bonus if the answer was exactly correct.

On top of this, participants were paid the lottery realization of the risk aversion elicitation stage. In total, an average participant earned 15.40 EUR.

C.1 Other figures

Information / Frame	Others thought: **"Men are better"**	Others thought: **"Women are better"**
"Math game"	$n_{Munich} = 230$ $n_{Bonn} = 73$	$n_{Munich} = 0$ $n_{Bonn} = 70$
"Memory game"	$n_{Munich} = 232$ $n_{Bonn} = 73$	$n_{Munich} = 0$ $n_{Bonn} = 71$

Figure 21: Treatment matrix

Figure 22: Timeline of the lab experiment

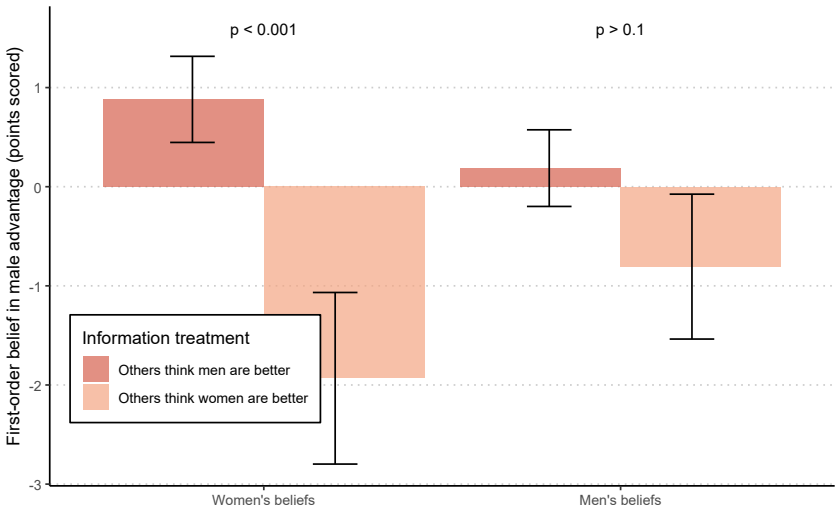


Figure 23: Information treatment effect on first-order beliefs

Error bars represent 95% confidence intervals.

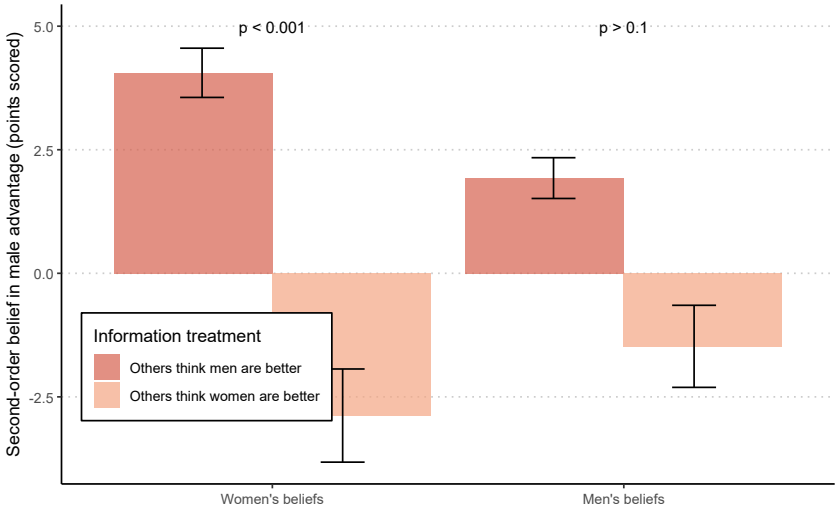


Figure 24: Information treatment effect on second-order beliefs

Error bars represent 95% confidence intervals.

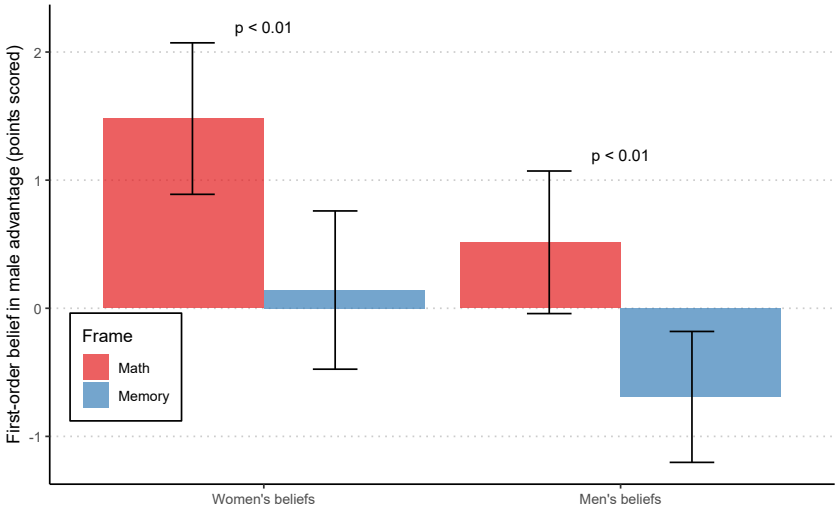


Figure 25: Framing treatment effect on first-order beliefs

Error bars represent 95% confidence intervals.

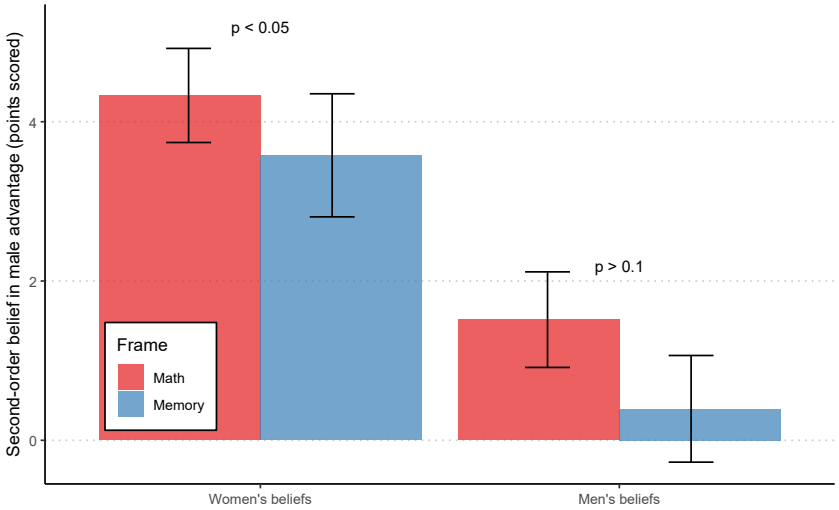


Figure 26: Information treatment effect on second-order beliefs

Error bars represent 95% confidence intervals.

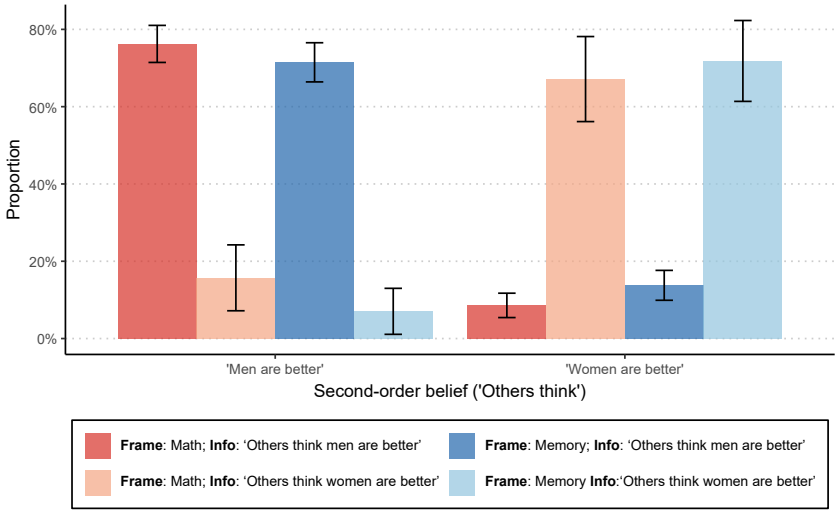


Figure 27: Treatment effect on first-order beliefs

Error bars represent 95% confidence intervals.

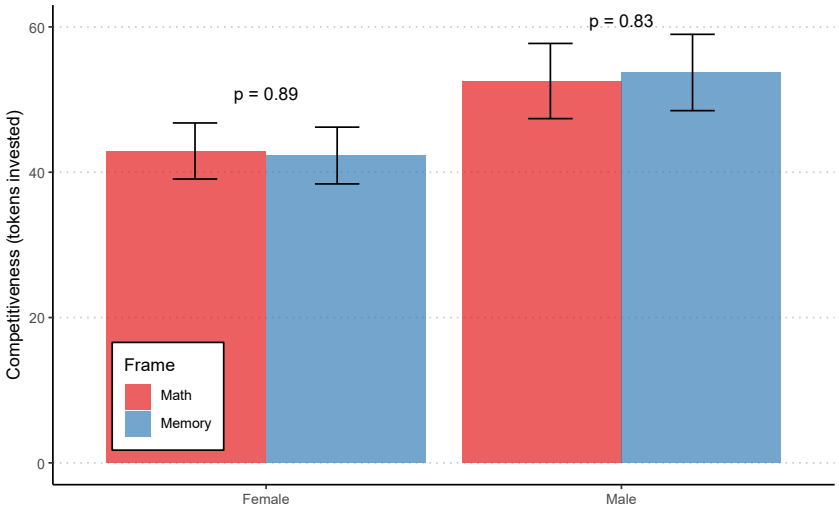


Figure 28: Framing effect on competitiveness (first stage)

Figure displays competitiveness of both genders split by gender and "Math" versus "Memory" framing. Competitiveness is measured by the number of tokens invested into the Tournament-rate in the first Tournament choice stage. Error bars represent 95% confidence intervals.

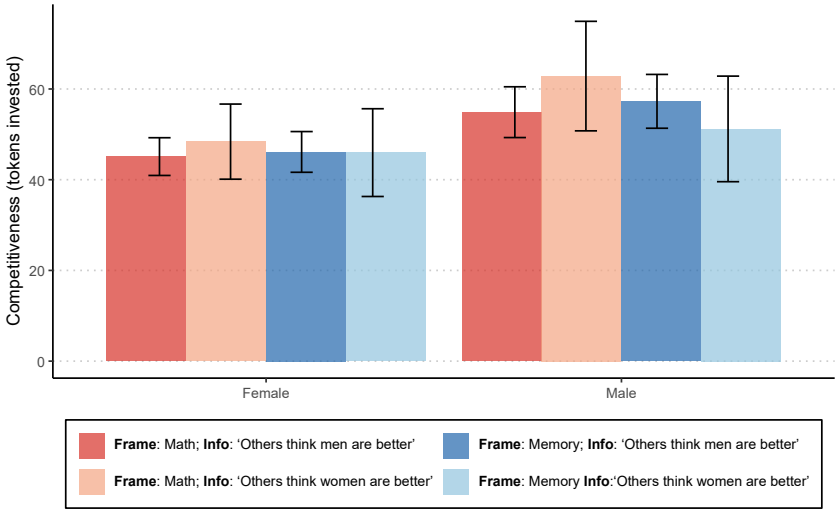


Figure 29: Treatment effect on competitiveness (second stage)

Figure displays competitiveness of both genders split by gender and the four treatments. Competitiveness is measured by the number of tokens invested into the Tournament-rate in the second Tournament choice stage. Error bars represent 95% confidence intervals.

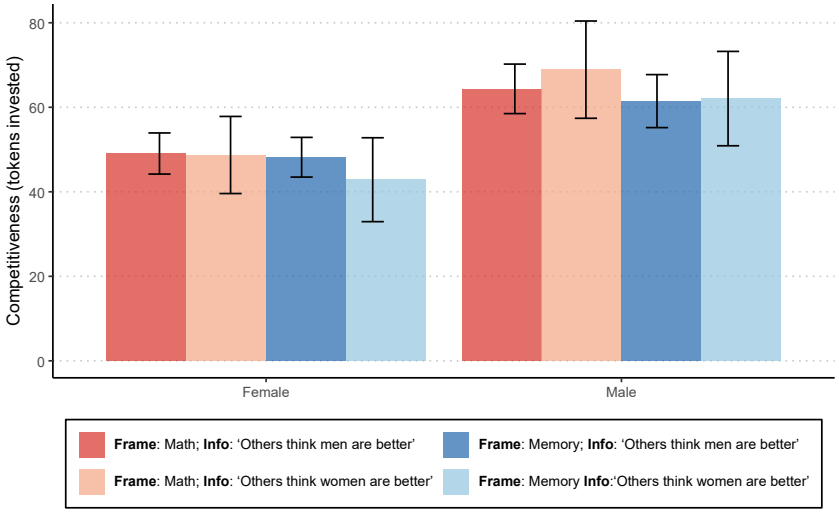


Figure 30: Treatment effect on competitiveness (third stage)

Figure displays competitiveness of both genders split by gender and the four treatments. Competitiveness is measured by the number of tokens invested into the Tournament-rate in the third Tournament choice stage. Error bars represent 95% confidence intervals.

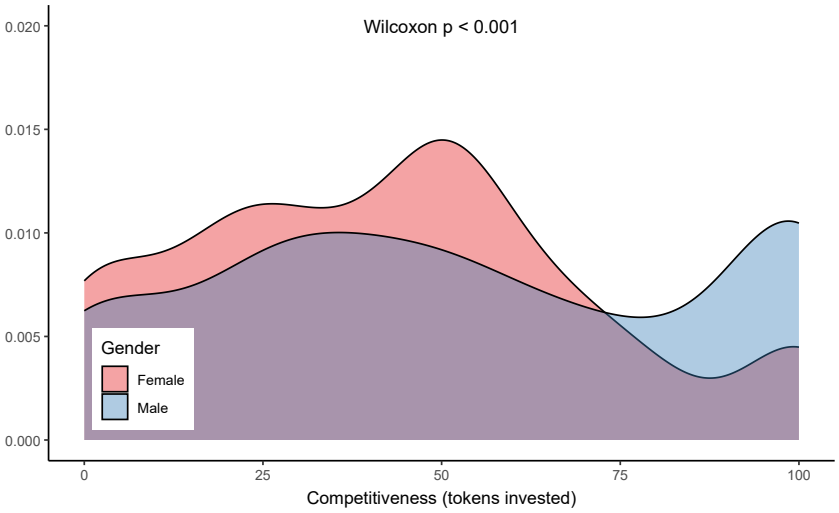


Figure 31: Distributional differences of competitiveness by gender (kernel density)

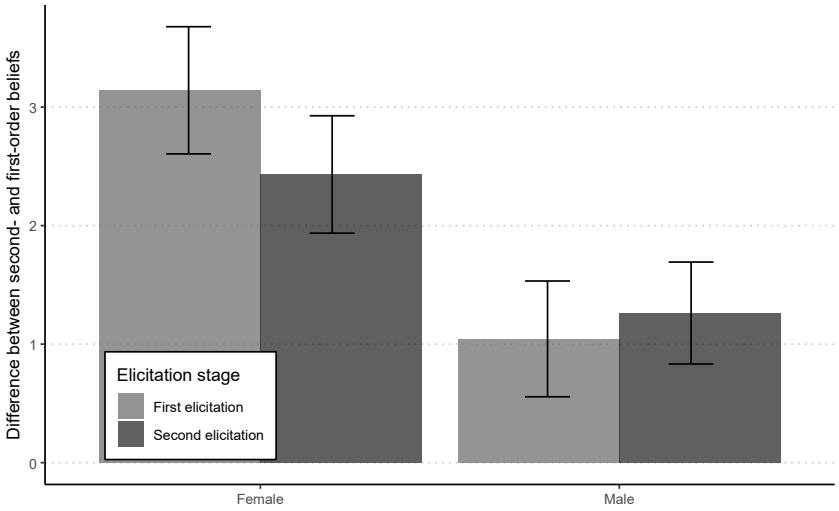


Figure 32: Difference between second- and first-order beliefs

Figure displays the average difference between second- and first-order beliefs per gender in two elicitation stages of the lab experiment, calculated as the participant's second-order belief in male advantage - participant's first-order belief in male advantage. Both male and female participants, on average, believe that others hold stronger beliefs in male advantage compared to their own first-order beliefs. Error bars represent 95% confidence intervals.

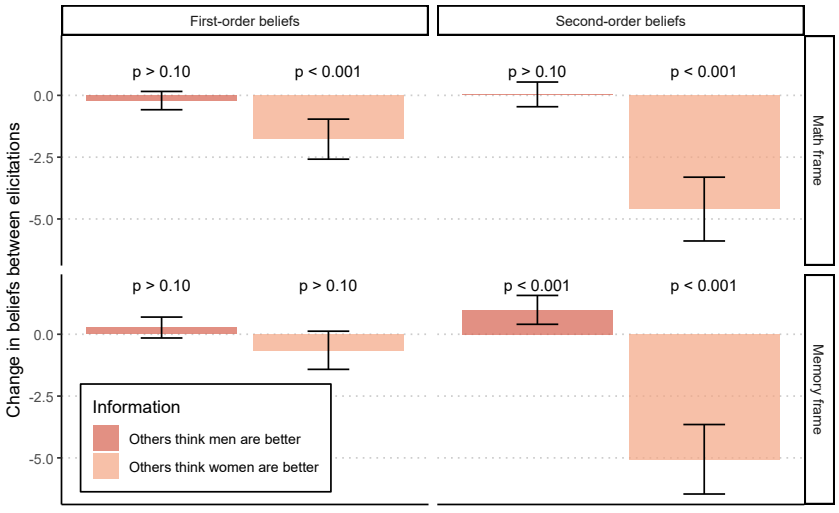


Figure 33: Within participant change in beliefs

Figure displays the effect of Information treatment on within participant change in beliefs within Math versus Memory frame. In each plot, the red (blue) bar represents the effect of informing participants that others think men (women) are better on the participant's belief. The left (right) column displays the treatment effect on first-order (second-order) beliefs. Error bars represent 95% confidence intervals.

C.2 Regression tables

Table 5: Balance table by lab

Characteristic	**Bonn** N = 287	**München** N = 462
Female	145 (51%)	265 (57%)
Age	24.7 (4.6)	24.7 (6.8)
Education		
Bachelors	83 (29%)	134 (29%)
Doctorate	2 (0.7%)	1 (0.2%)
GED (Hauptschulabschluss)	2 (0.7%)	1 (0.2%)
Havent graduated high school	2 (0.7%)	0 (0%)
High school graduate (Abitur)	161 (56%)	245 (53%)
Masters	29 (10%)	66 (14%)
Professional degree (Berufsexamen) - JD, MD, MBA	8 (2.8%)	15 (3.2%)
Employment		
Employed full-time	19 (6.6%)	22 (4.8%)
Employed part-time	14 (4.9%)	37 (8.0%)
Independent, or business owner	4 (1.4%)	8 (1.7%)
Out of labor force (e.g. retired or parent raising one or more children)	0 (0%)	1 (0.2%)
Out of work, or seeking work	11 (3.8%)	20 (4.3%)
Student	239 (83%)	374 (81%)
Income		
> 100.000 EUR	31 (11%)	88 (19%)
0 - 25.000 EUR	115 (40%)	146 (32%)
25.000 - 50.000 EUR	58 (20%)	93 (20%)
50.000 - 75.000 EUR	51 (18%)	91 (20%)
75.000 - 100.000 EUR	32 (11%)	44 (9.5%)

[1] n (%); Mean (SD)

Table 6: Summary statistics by gender and lab

Characteristic	**Female - Bonn** N = 145	**Female - München** N = 265	**Male - Bonn** N = 142	**Male - München** N = 197
Treatment				
Math frame & 'Men are better'	35 (24%)	133 (50%)	38 (27%)	97 (49%)
Math frame & 'Women are better'	38 (26%)	0 (0%)	32 (23%)	0 (0%)
Memory frame & 'Men are better'	36 (25%)	132 (50%)	37 (26%)	100 (51%)
Memory frame & 'Women are better'	36 (25%)	0 (0%)	35 (25%)	0 (0%)
Piece rate performance	24.1 (6.5)	22.2 (7.3)	24.9 (7.7)	23.1 (8.5)
Tournament performance	28.8 (6.7)	27.1 (8.3)	30.6 (7.8)	27.6 (10.0)
Competitiveness (1st stage)	43.9 (30.1)	41.9 (27.0)	56.4 (33.9)	50.8 (34.5)
Competitiveness (2nd stage)	46.2 (29.4)	45.8 (27.4)	57.0 (34.8)	55.7 (33.3)
Competitiveness (3rd stage)	48.0 (30.0)	48.2 (31.3)	64.3 (35.5)	62.7 (34.8)
Risk aversion switching point	470.0 (1,986.7)	505.6 (2,071.5)	476.4 (2,007.3)	357.0 (1,713.3)
Overconfidence	0.8 (1.9)	0.4 (1.7)	0.9 (1.8)	0.7 (1.6)
First-order beliefs (1st elicitation)	0.1 (4.0)	1.2 (4.6)	-0.2 (3.5)	0.0 (3.6)
Second-order beliefs (2nd elicitation)	4.1 (4.8)	3.9 (5.1)	0.8 (3.8)	1.0 (4.5)
First-order beliefs (1st elicitation)	-0.9 (4.0)	1.0 (4.1)	-0.3 (2.7)	0.2 (3.5)
Second-order beliefs (2nd elicitation)	0.7 (5.4)	4.0 (4.8)	0.1 (3.4)	2.1 (3.7)

[1] n (%); Mean (SD)

Table 7: Role of beliefs in competition gap

	1st stage	1st stage	2nd stage	2nd stage	3rd stage	3rd stage
Intercept	42.612***	43.142***	45.900***	46.162***	48.122***	48.558***
	(1.536)	(1.556)	(1.524)	(1.523)	(1.621)	(1.616)
Male	10.535***	10.094***	10.324***	10.073***	15.259***	14.832***
	(2.283)	(2.293)	(2.265)	(2.259)	(2.409)	(2.397)
F.O.B		−0.651		−0.701		−1.168**
		(0.346)		(0.367)		(0.390)
Male×F.O.B		1.561**		1.901**		2.164**
		(0.586)		(0.637)		(0.676)
Num.Obs.	749	749	749	749	749	749
R2	0.028	0.037	0.027	0.039	0.051	0.066

The table shows the results of a linear probability model (OLS). The dependent variable is competitiveness i.e., tokens invested to the Tournament-rate in the 3 Tournament choice stages. The first two columns refer to the first Tournament choice stage where participants actively competed against 5 others. The second two columns refer to the second Tournament choice stage where participants actively competed against the past scores of 5 others. The last two columns refer to the third Tournament choice stage where participants submitted their past score against the score of a randomly selected participant of the opposite gender. F.O.B (S.O.B) refer to participant's belief in male advantage calculated as belief in average male score minus belief in average female score. 1st Competition stage refers to the first Tournament choice stage where participants actively competed against 5 others. 2nd Competition stage refers to the second Tournament choice stage where participants actively competed against past scores of 5 others. 3rd Competition choice stage refers to the third Tournament choice stage where participants submitted their past score against the score of a randomly selected participant of the opposite gender. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8: Treatment effects on beliefs

	1st stage	1st stage	2nd stage	2nd stage
Intercept	1.048***	3.070***	0.997***	3.191***
	(0.209)	(0.253)	(0.210)	(0.241)
Memory Frame	−1.288***	−0.947**	−0.849**	−0.175
	(0.295)	(0.356)	(0.296)	(0.341)
Female congruent information			−2.411***	−5.077***
			(0.485)	(0.557)
Memory×Female congruent information			0.883	−0.474
			(0.683)	(0.786)
Num.Obs.	749	749	749	749
R2	0.025	0.009	0.053	0.198

The table shows the results of a linear probability model (OLS). The dependent variable is belief in male advantage calculated as belief in average male score minus belief in average female score. The first (last) two columns corresponds to the first (second) belief elicitation stage. F.O.B and S.O.B correspond to first- and second-order beliefs in male advantage calculated as the difference between first- or second-order belief in average male and female scores. Memory Frame refers to the dummy variable that equals one if the participant is assigned to the Memory framing treatment. Female congruent information refers to the dummy variable that equals one if the participant is assigned to receiving the information 'Others thought women perform better' compared to 'Others thought men perform better'. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 9: Competitiveness by treatments in all stages

	1st stage	1st stage	2nd stage	2nd stage	3rd stage	3rd stage	3rd stage
Intercept	42.612***	42.922***	45.900***	45.610***	48.122***	48.995***	48.619***
	(1.536)	(2.170)	(1.524)	(1.685)	(1.621)	(2.287)	(1.792)
Male	10.535***	9.629**	10.324***	10.482***	15.259***	16.238***	14.282***
	(2.283)	(3.243)	(2.265)	(2.520)	(2.409)	(3.418)	(2.679)
Memory Frame		−0.623				−1.755	
		(3.076)				(3.243)	
Male×Memory Frame		1.799				−1.897	
		(4.573)				(4.820)	
Female congruent information				1.606			−2.754
				(3.967)			(4.217)
Male×Female congruent information				−0.937			5.182
				(5.787)			(6.152)
Num.Obs.	749	749	749	749	749	749	749
R2	0.028	0.028	0.027	0.027	0.051	0.053	0.052

The table shows the results of a linear probability model (OLS). The dependent variable is competitiveness calculated as tokens invested to the Tournament-rate in the specific Tournament choice stage. The first two columns refer to the first Tournament choice stage where participants actively competed against 5 others. The second two columns refer to the second Tournament choice stage where participants actively competed against the past scores of 5 others. The last two columns refer to the third Tournament choice stage where participants submitted their past score against the score of a randomly selected participant of the opposite gender. Memory Frame refers to the dummy variable that equals one if the participant is assigned to the Memory framing treatment. Female congruent information refers to the dummy variable that equals one if the participant is assigned to receiving the information 'Others thought women perform better' compared to 'Others thought men perform better'. Performance refers to the score attained at the first round of the game i.e. the Piece rate round. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

C.3 Mediation analysis

Despite having no direct effects on competitiveness, treatments could still affect competitiveness indirectly through beliefs in male advantage. Although the original Baron and Kenny (1986) paper listed the presence of a significant total (direct) effect as a prerequisite for mediation, subsequent research, mostly from the experimental psychology literature, has shown that this condition is not necessary. In particular, Zhao et al. (2010) argue that mediation can exist even in the absence of a total effect, especially when opposing indirect effects cancel out. Indeed, there are conceptual reasons to believe our treatments could have opposing direct effects. Reminders of gender-related stereotypes may cause some individuals to withdraw from competition while provoking in others a motivation to fight the stereotype (Kray et al., 2001). Moreover, reactions to gender-related information can depend on broader political or self-image considerations, potentially leading to heterogeneous behavioral responses within the same treatment condition. Similarly, Shrout and Bolger (2002) emphasize that significant indirect effects can be present even when direct effects are absent, and their simulations demonstrate that designs with sufficient power can reliably detect such mediation. Given our sample size, our design has 83% power to detect indirect effects of 0.10 standard deviations or larger. As MacKinnon et al. (2007) note, the emphasis in mediation analysis should shift from total effects to directly testing the significance and magnitude of the indirect path.

There are two key assumptions for the validity of the average causal mediation effect (ACME) estimator (Imai et al., 2010): ignorability of treatment, $T(Y(t,m), M(t))X$ and ignorability of mediator, $MY(t,m)T, X$. The first is satisfied in a randomized experiment. The second assumption states that there should be no omitted variable that confounds the mediator-outcome pathway. The mediation analyses below are conducted using the "mediation" R-package by Tingley et al. (2014).

C.3.1 Average Causal Mediation Effects

The ACME estimators reported in the figures below are robust to the inclusion of key covariates that could confound beliefs and competitiveness: risk aversion, overconfidence, and performance on the task (Piece-rate).

Figure 34 displays the ACME estimators in the first and second competition entry stages. The ACME estimates on framing indicate that memory framing, relative to math framing, causally affects competitiveness through its impact on beliefs: It reduces tournament entry among men and increases it among women. The ACME estimates on the information treatment show that when participants are told others believe women perform better, compared to being told the opposite, beliefs are a causal channel. Specifically,

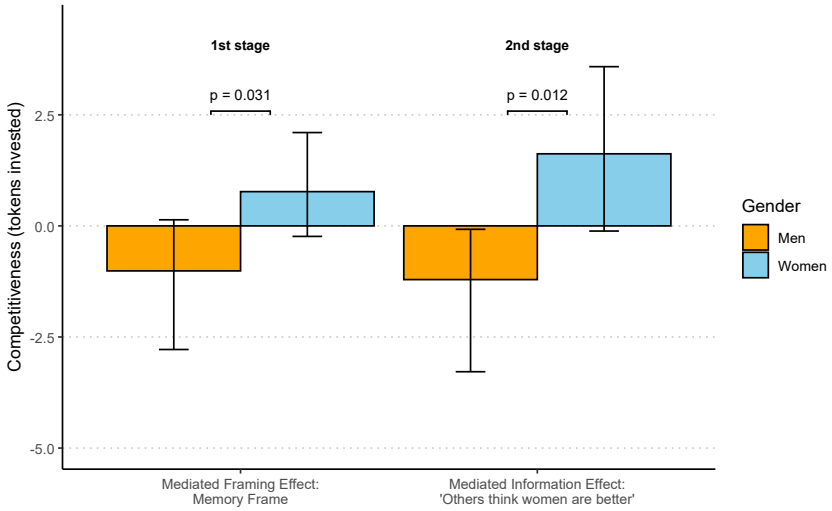


Figure 34: Treatment effects are mediated through beliefs in male advantage

Figure displays Average Causal Mediation Effects through beliefs in male advantage for both treatments in the first and second Tournament entry stages. Covariances include performance, overconfidence, and risk aversion.

male competitiveness decreases, while female competitiveness increases.

Figure 35 displays the ACME estimators in the third competition entry stage. Recall that in the third stage, participants submit their past performance to a Tournament against a participant of the opposite gender. The ACME estimates on framing and information tell comparable stories.

C.3.2 ACME are sensitive to potential omitted variables

The biggest pitfall in the mediation analysis is the existence of omitted variables that confound the mediator and the outcome. In our setting, it is easy to imagine such variables. For instance, it could very well be that women with stronger social desirability biases, upon receiving the information that others think men are better, would decrease their competitiveness and increase their belief in male advantage. This invalidates the estimator's claim to causality. However, what is important is how sensitive the results would be to such potential confounders. The sensitivity analysis reveals that the ACME estimators are, in fact, highly sensitive. For instance, Figure 36 displays the sensitivity analysis for the mediated information effect on women's competitiveness. The estimated mediation effect would disappear if there were an unobserved confounder that induces a negative correlation of just 0.2 between the errors in the mediator and outcome models. Another example of such an unobserved confounder would be internalized stereotypes. For example, it could be that men who have stronger internalized stereotypes would display stronger beliefs in male advantage and more competitive behavior. Overall, the sensitivity of the ACME estimates underscores that the evidence for a causal pathway through beliefs is suggestive but fragile and should be interpreted with caution.

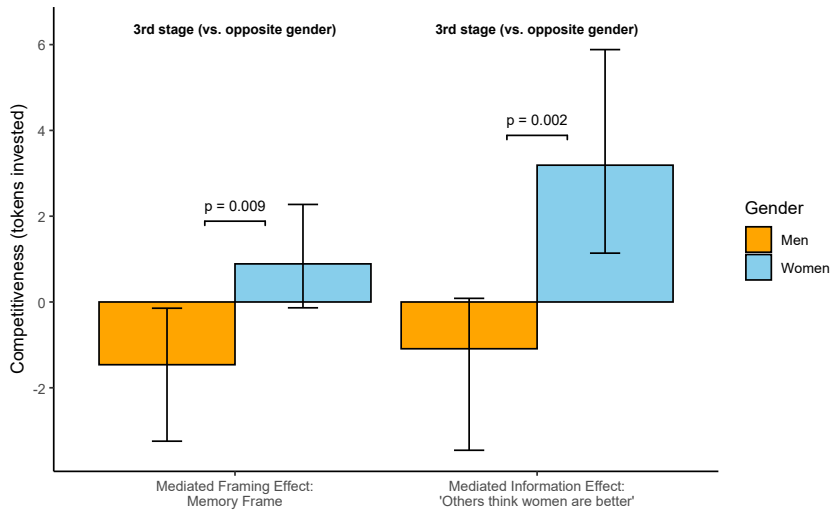


Figure 35: Treatment effects are mediated through beliefs in male advantage

Figure displays Average Causal Mediation Effects through beliefs in male advantage for both treatments in the final Tournament entry stage (against an opposite gender). Covariances include performance, overconfidence, and risk aversion.

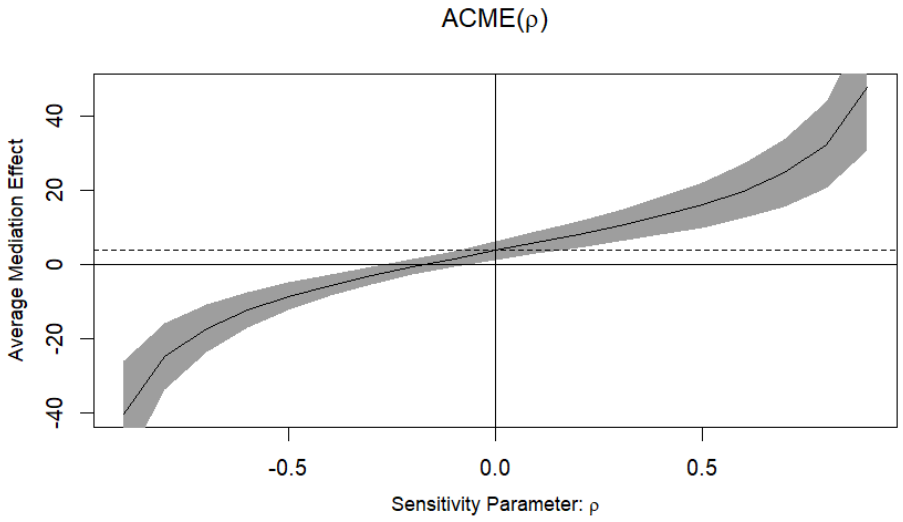


Figure 36: The ACME estimator is highly sensitive to potential omitted variables

D Online experiment

D.1 Procedures and timeline

After reading the instructions and completing comprehension check questions, participants played the Math-Memory task under the Piece-rate and the Tournament-rate schemes. Next, participants allocated 100 tokens between the competitive Tournament-rate and the noncompetitive Piece-rate schemes. Then, participants guessed the gender performance differences at the top and at the mean in a randomized order. Lastly, we implemented a binary measure of competitiveness by asking participants whether they would like to apply to their past score the Piece-rate or Tournament-rate scheme. At the end, participants completed a short demographic survey. The study included 3 attention checks and 3 comprehension questions, those who failed at least 2 attention checks as well as those who failed a comprehension question twice were dropped from the dataset. One random round was chosen to determine the bonus payment. If one of the belief elicitation rounds was chosen, participants earned a $10 bonus only if their guess was within 10% of the true value.

D.2 Additional figures and tables

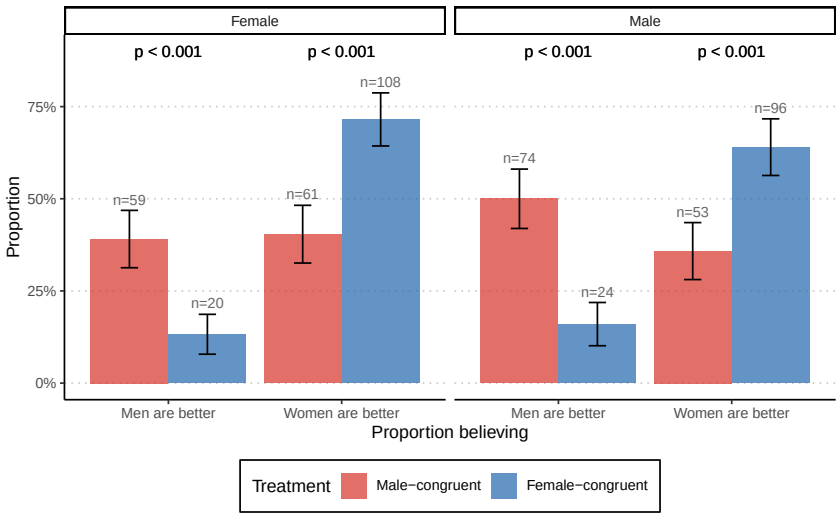


Figure 37: Treatment effect on beliefs about average performance (Online experiment)

Figure displays the proportion of participants believing in male versus female average performance advantage between treatments. Those in the "Men are better" ("Women are better") category believe that the average male (female) participant scored higher than the average female (male) participant in the Piece-rate round. The third belief category, i.e., those who believe men and women perform exactly the same is omitted for simplicity. Error bars represent 95% confidence intervals.

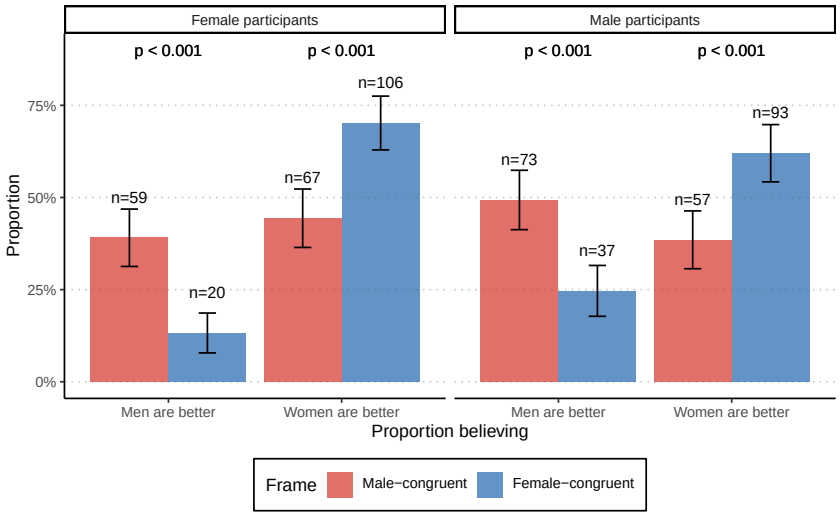


Figure 38: Treatment effect on beliefs about performance at the top (Online experiment)

Figure displays the proportion of participants believing in male versus female performance advantage at the top between treatments. Those in the "Men are better" ("Women are better") category believe that of the 600 participants, more men (women) reached the top 50 score in the Piece-rate round. The third belief category, i.e., those who believe men and women perform exactly the same is omitted for simplicity. Error bars represent 95% confidence intervals.

Table 10: Determinants of competitiveness (Online experiment)

	1st stage	1st stage	1st stage	2nd stage	2nd stage	2nd stage
Intercept	42.281***	42.274***	32.986***	−0.392***	−0.393***	−0.445
	(1.654)	(1.721)	(7.125)	(0.074)	(0.077)	(0.320)
Male	−1.194	−1.219	10.021	0.084	0.066	0.580
	(2.347)	(2.420)	(9.626)	(0.105)	(0.108)	(0.431)
Belief avg.		−0.005			−0.001	
		(0.298)			(0.013)	
Male×Belief avg.		−0.022			−0.014	
		(0.404)			(0.018)	
F.O.B top			0.398			0.002
			(0.297)			(0.013)
Male×Belief top			−0.477			−0.021
			(0.393)			(0.018)
Num.Obs.	600	600	600	600	600	600
R2	0.000	0.000	0.004			

The first two columns report OLS regressions where the dependent variable is competitiveness i.e., tokens invested to the Tournament-rate in the Tournament choice stage. The last two columns display probit regression results where the dependent variable is binary competitiveness i.e., whether the participant chose to apply the competitive Tournament-rate to their Piece-rate round performance. Belief avg. refer to participant's belief in male advantage calculated as belief in average male score minus belief in average female score. Belief top refer to participant's belief about how many of the 50 top performers in the Piece-rate round were men compared to women. Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

E Replication experiment

The aim of this experiment was to collect competition and performance data on four tasks: a quiz task, a Math-Memory task, a visual Memory task and a spot the difference task. In all tasks, participants have 120 seconds to maximize their score. Second, we wanted to validate that "math" vs. "memory" framing on the Math-Memory task affects the beliefs about male advantage.

The experiment was conducted on 16 April 2024 with 660 participants from Prolific's U.S. sample. The timeline of the experiment was as follows. After providing consent and basic demographic information, participants landed on the instructions page where we explained to them that they were in a group of 6 participants randomly chosen from the pool of US participants on Prolific. We explained that they would play two games and each game would have 3 rounds. Each of these rounds took 2 minutes. They were told that in addition to their completion payment, one of the 6 rounds would be randomly chosen to determine a bonus payment. The participants then played 2 minutes of the Math-Memory game in the Piece-rate setting. For half of the participants, the game was framed as the "Math game" and for the other half as the "Memory game." After the Piece-rate setting, participants played the game under the Tournament-rate scheme. Next, the computer chose one of the other three games randomly to be played by the participant and the participant played the game first in a Piece-rate setting and then in a Tournament-rate setting. After this, we elicited participants' competitiveness in the Math-Memory game by asking them to choose the Tournament-rate or the Piece-rate setting for their round 1 score. We elicited their competitiveness in the second game similarly.

The study included 2 attention checks and 3 comprehension check questions. Screenshots of the experiment interface can be found in the Online Appendix.

F Tasks

In this Appendix, we list the descriptions and examples for each of the tasks we have identified in the literature. Note that most of the pictures depicted in the following are screenshots from papers that have utilized the tasks. For some tasks such as the find hidden words and typing tasks, we have created the pictures from the descriptions to be used in the Belief Elicitation experiment. Some tasks do not have meaningful pictures to depict them.

F.1 Descriptions of Tasks

Math-Memory. We developed the Math-Memory task to create a task that lends itself naturally to both masculine and feminine framing. Solving the task requires participants to draw on both mathematical and memory-related skills, making it suitable for experimental manipulation of task stereotypes. In the paper, "Math task" ("Memory task) refers to the Math-Memory task under Math (Memory) framing.
Description. On the screen, the participants see a 4x4 matrix with covered cells. Behind each cell is a simple summation (e.g., 3 + 4) and a colored heart (red or black). To score a point, participants have to click on two cells at a time to find matching cells. Two cells are a match if the summations and the color of the hearts match. There are in total 4 matching cells in one deck of 16 cells. Once a participant completes a deck of 16 cells, another deck appears.

Visual Memory. On the screen, the participants see a 3x4 matrix with covered cells. Behind each cell is a picture of an animal. To score a point, participants have to click on two cells at a time to find matching cells. Two cells are a match if they hide the same picture of an animal. There are in total six matching cells in one deck of 12 cells. Once the participant completes a deck of 12 cells, another deck appears.

Adding numbers. Participants are shown a set of five two-digit numbers and must compute their sum in a limited time.

Anagram. Participants had to rearrange pairs of letters to form a meaningful word: TR, EA, TS, RE = RETREATS. LI, CU, NK, FF = CUFFLINK.

Ball bucket. In these tasks, participants were given 10 tennis balls that they had to throw into a small basket placed 10 feet away. For each successful throw they earned 10 points.

Count numbers. Participants count the number of ones in a 5x5 matrix. Once an answer is entered for one matrix, another is displayed. For each correct solution, they earn a point.

Data search. Participants were given a list of Fortune 500 companies and asked to scrape basic information (company name, revenues, profit, number of employees) and input their answers into a form in 5 minutes.

Emotion recognition. Participants see a series of faces and are asked to choose the correct emotion from four options. The images appear for two seconds; answers must be

submitted within 20 seconds.

Economics quiz. Participants answer multiple choice questions on microeconomics, earning one point per correct answer.

Find hidden words. Participants view a 10x10 letter matrix and find hidden words. Each matrix contains multiple solutions. Participants may skip to the next matrix at any time. One point is awarded per correct word.

Form words from letters. Participants are given eight letters and are asked to form as many words of at least three letters as they can within two minutes.

Letter difference. Participants view two 6x5 matrices filled with letters. While most letters are identical, two differ. The task is to click on the different letters. Each correct selection earns one point.

Maze. Participants navigate a maze using arrow keys on the keyboard to reach the goal point.

Mental rotation. A reference shape is shown alongside three candidates. One of the candidates is a rotated version of the reference shape. Participants must identify the correct one.

Multiplication
Participants are given two numbers and must multiply them. Each correct answer earns one point.

Number in numbers. A long 15-digit number and a short 3-digit number are shown. Participants use the digits of the long number to find summations that are equal to the shorter number.

Quiz. Participants answer multiple choice questions on various topics. Each question has four options and only one correct answer. One point per correct answer.

Rearrange words. Participants are given five words and must rearrange them to form a coherent sentence. Each correct sequence earns one point.

Search summation. A 3x3 grid of two-digit numbers is shown. The participants must find two numbers that sum to 100. Each correct pair earns one point.

Spot the difference. Two similar images are shown side-by-side. Participants must spot and mark up to 10 differences within a time limit.

Stock forecasting. Participants predict the price of a fictitious stock using two numeric cues. The price relationships, simple linear functions of the two cues, are not revealed. Participants learn from examples and feedback over 20 rounds.

Typing. Participants type five given letters into a text box. Each correct entry earns one point.

Verify arithmetics. Participants assess whether short arithmetic equations (sums or subtractions) are correct. Each correct verification earns one point.

Word-in-word. Participants are given a 12-letter word and must form as many valid words as possible from its letters. There are multiple correct answers.

Rope skipping. Participants have one minute to complete as many jumps as possible with the rope.

Toy fishing. Participants are sitting on a desk and catch a plastic "fish" using a magnetic pole then drop them into a bucket.

Sort shapes. Participants are asked to successfully order six eight-sided building blocks with various geometric shapes on each side from the smallest to the largest. Each side of a given block has one of six shapes.

Count letters. Participants count the number of appearances of a specific letter within a given set of 50 letters.

Lego building. Participants are given Lego bricks in specific shapes and colors and have to build as many rows as they can using specific instructions.

Count colors. Participants count the number of appearances of a specific color in a 5x5 matrix. Once an answer is entered for one matrix, another is displayed.

F.2 Pictures of some tasks

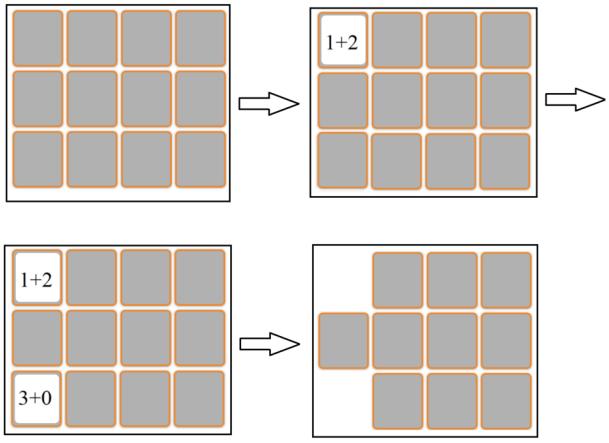


Figure 39: Math-Memory

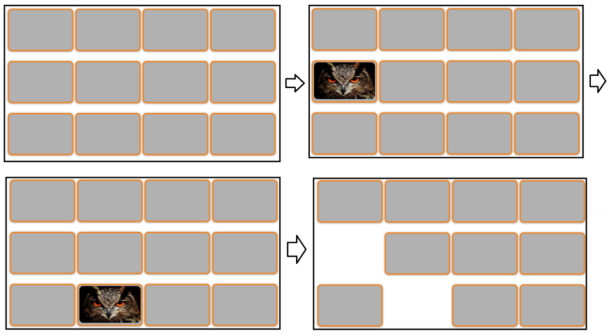


Figure 40: Visual Memory

21	35	48	29	83	

Figure 41: Adding numbers

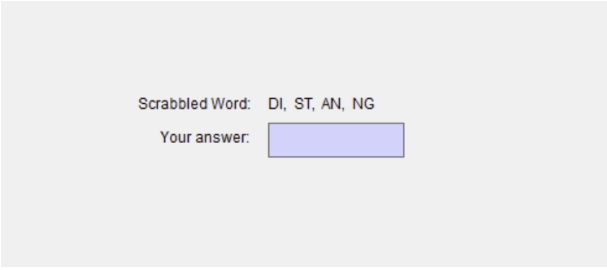


Figure 42: Anagram

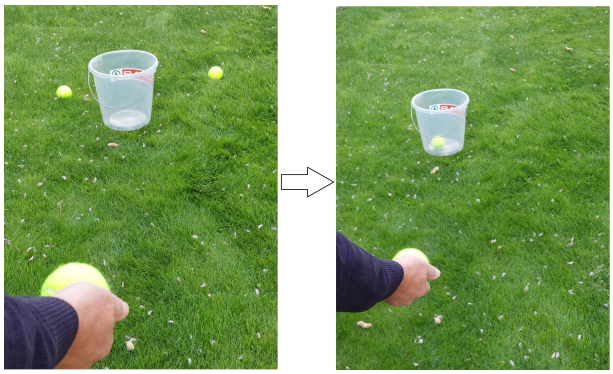


Figure 43: Ball bucket

0	0	1	1	0
0	0	1	0	1
0	1	1	0	1
1	0	1	1	0
1	0	0	0	1

Figure 44: Count numbers

3. **This face is expressing...**

- ○ Sadness
- ○ Pain
- ○ Anger
- ○ Disgust

Figure 45: Emotion recognition

G	W	W	S	C	W	O	R	L	D
C	U	J	T	N	M	O	O	N	I
O	X	H	A	P	Y	H	U	W	E
D	S	S	R	Y	O	E	P	L	A
E	P	R	G	T	T	L	S	R	R
J	A	R	R	H	V	L	U	Y	T
B	C	B	I	O	Y	O	S	V	H
X	E	J	D	N	S	U	N	P	C
W	Z	B	P	H	B	R	Y	P	G
U	M	F	M	K	L	T	J	U	Z

An example task (solutions in color)

Figure 46: Find hidden words

LEFT MATRIX

Y	K	C	M	F
C	M	E	Y	C
X	G	S	X	M
F	D	K	V	T
I	A	Z	D	Z
U	V	A	G	C

RIGHT MATRIX

Y	K	H	M	F
C	M	E	Y	C
X	G	S	X	M
F	D	K	V	T
I	A	Z	M	Z
U	V	A	G	C

Figure 47: Letter difference

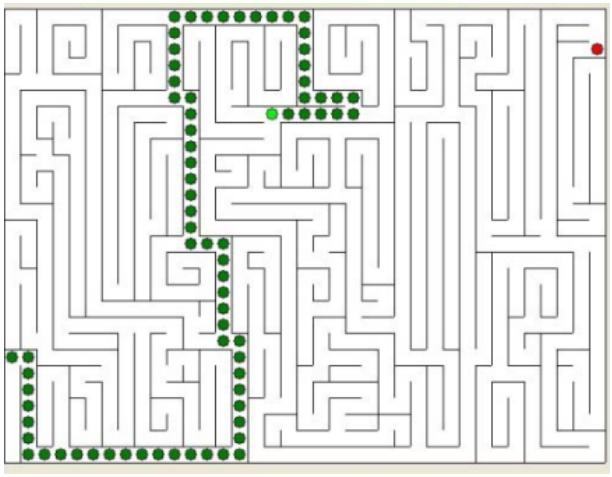


Figure 48: Maze

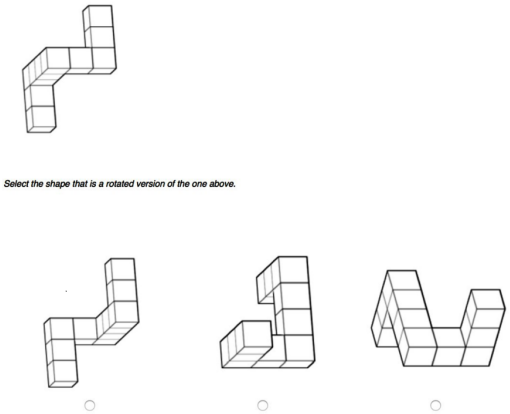


Figure 49: Mental rotation

1. 7 x 8 = ?
2. 9 x 6 = ?
3. 4 x 11 = ?
4. 12 x 3 = ?
5. 5 x 5 = ?
6. 8 x 2 = ?
7. 3 x 7 = ?
8. 11 x 4 = ?
9. 6 x 9 = ?
10. 10 x 12 = ?

Figure 50: Multiplication

Puzzle sequence: 436771974115604

Target number: 135

Figure 51: Number in numbers

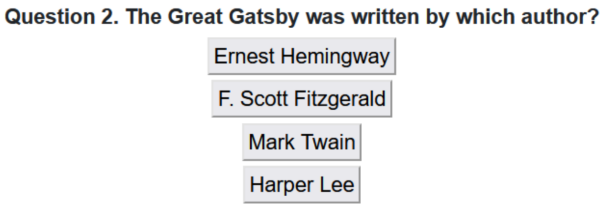


Figure 52: Quiz

Word 1	*Word 2*	*Word 3*	*Word 4*	*Word 5*
weather	fine	is	The	today

Figure 53: Rearrange words

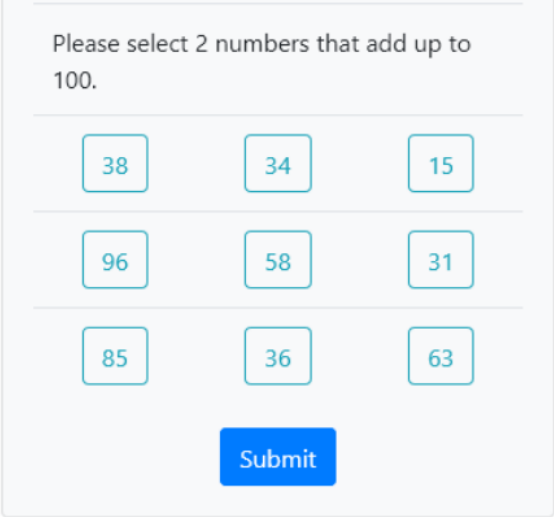


Figure 54: Search summation

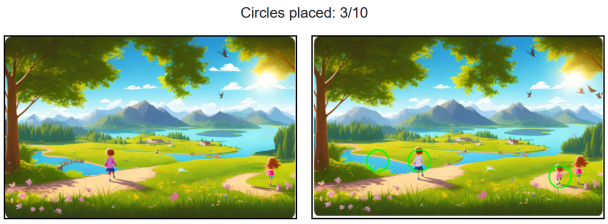
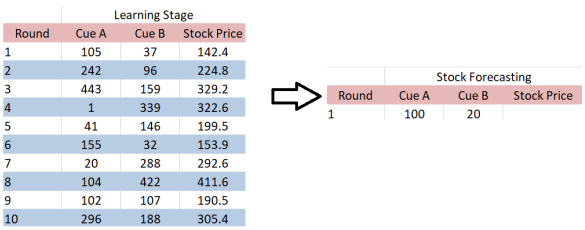


Figure 55: Spot the difference



Learning Stage

Round	Cue A	Cue B	Stock Price
1	105	37	142.4
2	242	96	224.8
3	443	159	329.2
4	1	339	322.6
5	41	146	199.5
6	155	32	153.9
7	20	288	292.6
8	104	422	411.6
9	102	107	190.5
10	296	188	305.4

Stock Forecasting

Round	Cue A	Cue B	Stock Price
1	100	20	

Figure 56: Stock forecasting

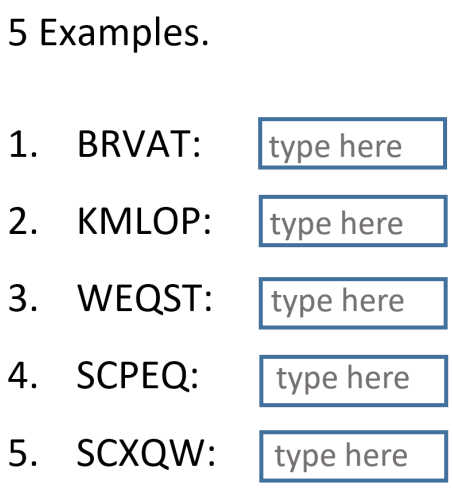


Figure 57: Typing

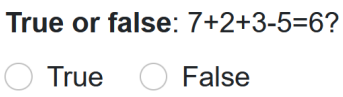


Figure 58: Verify arithmetics

Figure 59: Word-in-word