Gender Differences in Performance Under Competition:

Maastricht University

Diogo Geraldes, Arno Riedl,
Martin Strobel

Gender Differences in Performance under Competition: Is there a Stereotype Threat Shadow?

RM/21/002

ISSN: 2666-8807

GSBE

Maastricht University School of Business and Economics
Graduate School of Business and Economics

P.O Box 616
NL- 6200 MD Maastricht
The Netherlands

Gender Differences in Performance Under Competition: Is There a Stereotype Threat Shadow?*

Diogo Geraldes Arno Riedl Martin Strobel

December 23, 2020

Abstract

The gender gap in income and leadership positions in many domains of our society is an undisputed pervasive phenomenon. One explanation for the disadvantaged position of women put forward in the economic and psychology literature is the weaker response of women to competitive incentives. Despite the large amount of literature trying to explain this fact, the precise mechanisms behind the gender difference in competitive responsiveness are still not fully uncovered. In this paper, we use laboratory experiments to study the potential role of stereotype threat on the response of men and women to competitive incentives in mixed-gender competition. We use a real effort math task to induce an implicit stereotype threat against women in one treatment. In additional treatments we, respectively, reinforce this stereotype threat and induce a stereotype threat against men. In contrast to much of the literature we do not observe that women are less competitive than men, neither when there is an implicit nor when there is an explicit stereotype threat against women. We attribute this to two factors which differentiates our experiment from previous ones. We control, first, for inter-individual performance differences using a within-subject design, and, second, for risk differences between non-competitive and competitive environments by making the former risky. We do find an adverse stereotype threat effect on the performance of men when there is an explicit stereotype threat against them. In that case any positive performance effect of competition is nullified by the stereotype threat. Overall, our results indicate that a stereotype threat has negative competitive performance effects only if there is information contradicting an existing stereotype. This suggests that the appropriate intervention to prevent the adverse effect of stereotype threat in performance is to avoid any information referring to the stereotype.

JEL Classification: C91, D01, J16

Keywords: Competitiveness, gender gaps, stereotype threat, experiment

*Diogo Geraldes: Utrecht School of Economics, Utrecht University, the Netherlands, email: diogogeraldes@gmail.com; Arno Riedl (corresponding author): CESifo, IZA, Netspar and Maastricht University, Department of Micoreconomics and Public Economics, the Netherlands, email: a.riedl@maastrichtuniversty.nl; Martin Strobel: Department of Microeconomics and Public Economics, Maastricht University, the Netherlands, email: m.strobel@maastrichtuniversty.nl. We would like to thank the participants in the Experimental design seminar series at Maastricht University, International ESA conference, TIBER Symposium at Tilburg University, EEA-ESEM congress, and SITE workshop at Stanford University for invaluable feedback. We are particularly indebted for the comments of Simon Gächter and Ernesto Reuben. We are grateful that Thomas Dohmen provided us with his z-Tree program for the real-effort task.

1 Introduction

The existence of a gender gap in wages and positions in the labor market is an undisputed stylized fact. In 2018, women earned, on average, 14.8% less than men in the European Union (Eurostat, 2020). Further, in academics for instance, only 14% of the heads of institutions in the European Union with the capacity to deliver a Ph.D. are women (Directorate–General for Research and Innovation (European Commission), 2019). At CEO positions among top businesses, the figures for women are even more somber.[1]

One common explanation for these gender differences in the workplace is discrimination (e.g., Black & Strahan, 2001). Other explanations are women's higher sensitivity to work-family conflicts and women's weaker negotiation behavior (e.g., Babcock & Laschever, 2003). Another important explanation that emerged from the psychological and experimental economics literature is that women are less responsive to competitive incentives. However, the precise mechanisms behind this gender difference are still not fully known. In this paper, we examine one such potential mechanism by studying the impact of gender stereotypes on men's and women's performance under competition.[2]

A bulk of studies has shown that when men and women have to compete against each other in stereotypically regarded male-dominated contexts, men generally improve their performance under competition whereas this is not (or much less) the case for women (e.g., Gneezy, Niederle, & Rustichini, 2003; Gneezy & Rustichini, 2004; Günther, Ekinci, Schwieren, & Strobel, 2010; Shurchkov, 2012; Iriberri & Rey-Biel, 2017; Iriberri & Rey-Biel, 2019).[3] The psychology literature proposes 'Stereotype Threat' as a possible cause for the observed gender difference (Steele & Aronson, 1995). Stereotype threat is essentially a situational predicament where a member of a group is apprehensive by the possibility of confirming a negative stereotype about their group. Psychology research on stereotype threat has shown that non-competitive performance of a targeted group is harmed when a stereotype threat is present.[4] In our paper we connect the stereotype threat idea to the gender gap under competitive incentives.

[1]For an overview of gender differences in the labor market, see, e.g., Blau, Ferber, and Winkler (2013).

[2]The body of research on gender competition also addresses the important—but distinct—question regarding gender differences in self-selection into competitive environments (see, e.g., the seminal work of Niederle & Vesterlund, 2007, more recently, Buser, Niederle, & Oosterbeek, 2014; van Dolder, van den Assem, & Buser, 2020, and in relation to stereotypes and self-selection Hernandez-Arenaz, 2020, and the references therein).

[3]A notable exception is Dreber, Von Essen, and Ranehill (2011) who, in an experiment with Swedish children, do not find gender difference in competitive performance in a running task.

[4]For extensive reviews of the stereotype threat literature in psychology, see Inzlicht and Schmader (2012) and Spencer, Logel, and Davies (2016).

Specifically, we conduct a controlled laboratory experiment to investigate men's and women's performance in a mathematical real-effort task. In various treatments we examine whether confirming and disconfirming , implicit and explicit, stereotypical information affects the performance of men and women in mixed-gender competition.

To our knowledge, only two studies have hitherto studied stereotype threat in an environment where men and women have to compete with each other. Günther et al. (2010) find that women perform worse than men under competitive incentives for a task stereotypically favoring men, but perform at least as well as men under competition for a gender-neutral task and a task stereotypically favoring women. The authors suggest that women compete less in tasks for which men are stereotypically believed to perform better than women because of stereotype threat. However, with their design they can not show that this link actually exists.

Iriberri and Rey-Biel (2017) also find that women underperform under competition only when a task is perceived as male-dominated. Their study offers an insightful investigation into the effect of induced beliefs about the opponent's actual ability on women's competitive performance. Specifically, in all their treatments, it is publicly known that competing pairs are matched according to baseline performance (i.e., it is known to participants that each one will compete against another one with a similar ability to perform the real-effort-task).[5] We note, that this design only captures a specific aspect of stereotype threat because stereotypes relate to widely held beliefs regarding the comparison between two groups. In the context of a competition, assessing the effect of stereotype threat on competitive performance requires assessing how induced beliefs about the comparison of two specific groups affects the competitive performance of the negatively stereotyped person against an opponent of the outgroup.

According to the psychological literature, stereotype threat experience could be either implicitly or explicitly activated (Smith & White, 2002). Implicit activation of stereotype threat refers to situations where simply placing the targeted individual in a context where the associated stereotype is well known, but not explicitly highlighted, is sufficient to induce the targeted individual to experience the threat. Explicit activation of stereotype threat refers to situations where the targeted individual experiences the threat by explicitly confronting the individual with the stereotype. Existing evidence suggests that implicitly in-

[5]Moreover, in the treatments where the authors aim to enhance stereotype threat against women by providing information about ability differences, they do so by informing competing pairs about the relative baseline performance of the about-to-be opponent. In other words, the authors assess how induced beliefs about the actual ability of the opponent affects women's competitive performance.

duced stereotype-based expectations may be important (Günther et al., 2010). However, in the field, women are often (made) explicitly aware of that positions are dominated by men or that people expect men to be more successful than women in those domains (e.g., women competing for academic positions in math-intensive fields or aiming for top-paid corporate positions). Therefore, we consider it as important to also study gender competition in situations where stereotype-based expectations are explicitly induced.

The most prominent effect of stereotype threat reported in the psychology literature is that it undermines task performance of various groups across multiple domains. For example, this effect has been shown among African-Americans (see, e.g., Steele & Aronson, 1995) and Latinos (see, e.g., Schmader & Johns, 2003) when compared to Caucasians on tests labeled as indicators of intellectual ability. It has also been observed among women when compared to men during tests evaluating mathematical ability (see, e.g., Spencer, Steele, & Quinn, 1999), and among Caucasian men in math-tests when they were informed about Asian-Americans' superior ability in mathematics (Aronson et al., 1999). Essentially, stereotype threat theory predicts that an existing performance gap between two groups should be larger in a situation where one of the groups is subject to stereotype threat, in comparison to a situation where no stereotype threat exists. Importantly, this psychology research on stereotype threat focuses on non-competitive situations.

In a nutshell, in our experiment, we compare within-subject how men's and women's performance in a math task changes from a performance stage without competition to a performance stage with mixed-gender competition. We conduct four between-subjects treatments varying the competitiveness of the second stage and the stereotype threat at the second stage, respectively. By using a math task, but giving no additional information, we induce an implicit stereotype threat against women in one treatment. In another treatment we reinforce this stereotype by providing participants with truthful information about male superiority in a subset of mathematical problems. In a third treatment, we induce a stereotype threat against men by informing participants truthfully of the superiority of women in another subset of mathematical problems. In a fourth (control) treatment both performance stages are non-competitive.

We hypothesize that there are two opposing effects regarding competitive performance in our experiment. First, for the gender that is not the target of a stereotype threat, we expect that competition induces participants to exert higher effort at the competitive performance stage, relative to the non-competitive performance stage. Second, for the gender that is the target of a stereotype threat, we expect that the threat counteracts the positive

effect of competition on performance. Thus, we expect that the competitive performance of participants depends on whether or not they are part of the target group of the stereotype threat.

In brief, our results can be summarized as follows. First, we do not replicate the finding that women's performance is impaired in situations entailing a mathematical stereotype threat against women. When we induce, either implicitly or explicitly, a mathematical stereotype threat against women, we observe that women improve performance under competition as much as men do. However, we do find evidence for an adverse stereotype threat effect on the performance of men when we induce an explicit mathematical stereotype threat against them. Specifically, in this case men do not improve their performance under competitive incentives. Moreover, we provide evidence that men's relative underperformance is due to both fewer attempts to solve problems and decreased accuracy when attempting to solve a problem. This is consistent with the idea that performance is impaired by a stereotype threat shadow.

The remainder of the paper is organized as follows. In Section 2, we describe the design of the lab experiment followed by Section 3, where we state our hypotheses. In Section 4 we present our results, which includes hypotheses testing and an additional exploratory analysis. Section 5 concludes the paper.

2 Design of the experiment

We implement a combined between- and within-subjects design. Between-subjects, we conduct a baseline treatment without competition, and three treatments where participants have to compete for a prize. Across the three competitive treatments, we vary the stereotype threat which we describe in detail below. Importantly, each treatment consists of two stages. In all treatments, in the first stage, participants perform individually a real-effort task under a non-competitive payment scheme. In the second stage, we randomly pair each participant with a participant of the opposite gender to subsequently again perform an individual real-effort task. In the baseline treatment, the payment scheme at the second stage is non-competitive as at the first stage. In the three other treatments, participants perform under a competitive payment scheme at the second stage. Our main variable of interest is the within-subject change in performance between stages. This allows us to analyze the

performance in the real-effort task while controlling for individual differences, including participants' task productivity.[6]

The real-effort task consists of multiplying one- and two-digit numbers and, following Dohmen and Falk (2011), we implement five different levels of difficulty.[7] The problems are presented on computer screens and participants have to type each answer into a box and confirm it by clicking an "OK"-button with the mouse. Each time an answer is submitted, participants are informed whether the answer is correct. If the answer is correct, a new problem instantaneously appears on the screen. If the answer is wrong, the same problem has to be tackled until the correct solution is provided. This procedure helps to prevent participants from guessing. Participants are not allowed any aid (calculator, paper and pencil, etc.) to solve the multiplication problems.

We employ a mathematical task for the following reasons. First, to assess the potential effects of a stereotype threat, we need a task that is stereotyped. The multiplication task is suitable because it is well-documented that there is the stereotype that men are better at mathematics (e.g., Spencer et al., 1999).[8] Second, a crucial feature of our design is the manipulation of the stereotype threat. The multiplication task is ideal for this aspect, because there is reliable information concerning gender differences in mathematical ability that we can use to induce both confirming and disconfirming stereotype threats without deceiving participants. We discuss this element in more detail below.

At the beginning of the experiment, participants are informed that it consists of two performance stages, and that they receive specific instructions before the start of each stage. Participants are further informed that they can earn money in each performance stage, and that the possible earnings in the first stage are independent of their own and others' behavior in the second stage. Also, they do not receive feedback on performance at the end of each stage, neither in absolute terms nor relative to others.[9] Next, we describe in detail the sequence of events in the experiment.

Multiplication task practice. The experiment starts with a non-incentivized practice task in which participants are asked to solve as many multiplication problems as possible

[6] Gneezy and Rustichini (2004) used this method in a field experiment in which they study gender differences in children's performance under competition. Iriberri and Rey-Biel (2017) also use a within-subject measure of performance.

[7] Examples of the five levels are: 11 x 9 (level 1); 3 x 32 (level 2); 6 x 43 (level 3); 4 x 68 (level 4); 7 x 89 (level 5).

[8] In a separate online questionnaire, we find evidence that both men and women from our pool of participants indeed hold this stereotype. We present the results of the online questionnaire in Section 3.

[9] The full set of instructions used in the experiment can be found in the Web Appendix.

within 2 minutes. This task is intended to familiarize participants with the real-effort task. Participants are informed that it is in their best interest to gain practice because they can earn money with the multiplication task in the performance stages.

First stage performance. In this stage, we elicit participants' performances under non-competitive monetary incentives. The performance in this stage serves as the baseline for comparison with the performance in the subsequent competitive second stage. Before participants start with the multiplication task, they are informed that they have been randomly paired with another participant present in the lab, without making any reference to gender. Participants perform the multiplication task for 5 minutes under a random pay incentive scheme. Specifically, participants are informed that at the end of the experiment one participant in each pair will be chosen with equal probability to be paid out. If chosen, a participant earns €0.40 for each correctly solved multiplication problem, otherwise the participant earns nothing. We implement random payment in the non-competitive environment to control for the element of risk, which is inevitably present in the competitive environments.[10]

Confidence level elicitation. Immediately after the first stage, we ask each participant to estimate their relative performance compared to four randomly chosen participants present in the lab. Specifically, participants are asked to indicate their best estimates (in percentage) that exactly 0, 1, 2, 3 or 4 of the randomly chosen other participants correctly answered more problems in the first stage. We monetarily incentivize this belief elicitation using a quadratic scoring rule (Offerman, 1997) (for details, see the Web Appendix).

Second stage performance. In this stage, we distinguish four environments: a non-competitive (control) environment and three environments with the same competitive incentives. In each competitive environment, we randomly form mixed-gender pairs where men and women compete with each other, and make participants aware of it by mentioning in the instructions that they are paired with an opposite gender participant. Subsequently, participants perform the multiplication task for 5 minutes under a winner-takes-all tournament incentive scheme. That is, in each pair, the participant who correctly solves more multiplication problems earns €0.40 for each solved multiplication. The other earns nothing. In case of a tie, both participants earn €0.20 for each correctly solved multiplication. All participants face the same sequence of multiplication problems.

[10] If a standard piece-rate were used as the non-competitive payment scheme, gender differences in attitude towards risk (Croson & Gneezy, 2009) may obscure the pure competition effect on performance.

Next we describe how we induced different stereotype threats which constitutes our main treatment variable and which is the only dimension in which the three competitive environments differ.

Implicit Stereotype Threat Against Women (IST-W). In this treatment, we merely provide the instructions for the multiplication task. According to the idea of an implicit activation of stereotype threat (e.g., Smith & White, 2002), the use of a mathematical task in our experimental setting should be a sufficient situational clue to induce women to experience stereotype threat because the known mixed-gender pair formation makes gender salient, i.e., it makes women aware that they have to compete with a men.

Explicit Stereotype Threat Against Women (EST-W). In this treatment, we reinforce the stereotype threat against women by explicitly inducing a negative stereotype about women's ability to perform a mathematical task. Just before starting to compete, each participant unexpectedly faces the following information during 40 seconds:

"Before starting the multiplication task in this stage, please read carefully the following information:

In order to assess the magnitude of gender differences in mathematics performance, three leading researchers, J. Hyde, E. Fennema and J. Lamon, performed an evaluation of 100 studies in this field. In their paper, we can read the following: "refined discussions generally conclude that the overall differences in mathematics performance (. . .) appear in adolescence and usually favor boys in tasks involving problem solving."

In "Gender Differences in Mathematics Performance: A Meta-Analysis," Psychological Bulletin.

If you wish, you can inspect this paper after the experiment."

Explicit Stereotype Threat Against Men (EST-M). In this treatment, we contradict the stereotype that men are better at mathematics by explicitly inducing a negative stereotype about men's ability to perform a mathematical task. Just before starting to compete, each participant unexpectedly faces the following information during 40 seconds:

"Before starting the multiplication task in this stage, please read carefully the following information:

In order to assess the magnitude of gender differences in mathematics performance, three leading researchers, J. Hyde, E. Fennema and J. Lamon, performed an evaluation of 100 studies in this field. In their paper, we can read the following: "refined discussions generally

conclude that the overall differences in mathematics performance (. . .) appear in adolescence and usually favor girls in tasks of computation."

In "Gender Differences in Mathematics Performance: A Meta-Analysis," Psychological Bulletin.

If you wish, you can inspect this paper after the experiment."

We note that the quotes are not perfectly symmetric because we needed to avoid deceiving participants. With these quotes, extracted from Hyde, Fennema, and Lamon (1990) (p. 140), we use truthful information that allows us to implement the desired manipulations while still minimizing text asymmetry.

Finally, we conduct a control treatment to assess whether there are any significant learning or fatigue effects during the performance of the multiplication task.

Twice Random Pay (TRP). In the second stage of this treatment, participants perform the multiplication task for 5 minutes under the random pay incentive scheme, as in the first stage.

In each treatment, after the end of the second performance stage, we elicit participants' attitudes towards risk and competition.

Risk attitude elicitation. We elicit each participant's response to the Dohmen et al. (2011) general risk question: "How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick a box on the scale, in which the value 0 means 'not at all willing to take risks' and the value 10 means 'very willing to take risks'."[11]

Competitive attitude elicitation. We elicit each participant's response to the Machiavelli personality test, also known as the Mach IV test, in which high scores reliably predict competitive behavior (Christie & Geis, 2013) (for details, see the Web Appendix).

2.1 Experimental procedure

Participants in the experiment were recruited from the participants pool of the Behavioral and Experimental Economics Laboratory (BEELab) at Maastricht University, using the ORSEE software (Greiner, 2015). The experiment was fully computerized using the z-Tree software (Fischbacher, 2007) and conducted in the BEELab. We conducted eight sessions (two sessions with each of the four treatments). In total, 192 participants took part in

[11] In a paid field experiment, Dohmen et al. (2011) show that responses to this question reliably predict incentivized lottery choices.

the experiment (40, 46, 50 and 56 in the IST-W, EST-W, EST-M, and TRP, respectively). By design, an equal number of men and women participated in each session.[12] The participants were mostly students of business and economics at Maastricht University with an average age of 22.[13] Each session lasted around 90 minutes and average earnings were €16.30. All decisions and payments were treated confidentially.

3 Hypotheses

Our main variable of interest is the change in performance from the non-competitive first stage to the (non-)competitive second stage. Table 1 summarizes the incentives and stereotype threat present in each stage for each treatment.

Table 1. Payment incentive and stereotype threat per stage

	1st stage		2nd stage	
Treatment	Incentive	Stereotype Threat	Incentive	Stereotype Threat
IST-W	Random Pay	None	Competition	Implicit Against Women
EST-W	Random Pay	None	Competition	Explicit Against Women
EST-M	Random Pay	None	Competition	Explicit Against Men
TRP	Random Pay	None	Random Pay	None

Note: IST-W ... Implicit Stereotype Threat Against Women; EST-W ... Explicit Stereotype Threat Against Women; EST-M ... Explicit Stereotype Threat Against Men; TRP ... Twice Random Pay.

The control treatment consists of two identical stages in which participants perform under a non-competitive payment scheme. Based on Dohmen and Falk (2011), we do not expect any significant learning or fatigue effects during the performance of the multiplication task.

Hypothesis 1. *In treatment TRP, both men's and women's change in performance between stages is not significant.*

Previous research has shown that competitive incentives generally increase performance relative to a non-competitive environment (e.g., Gneezy et al., 2003). Therefore, we expect that the winner-takes-all tournament payment scheme in the three treatments with competition will lead to an increase of participants' performance, especially for those who are not

[12] See the Web Appendix for the procedure to ensure a perfectly balanced gender composition in each session.

[13] 86% of the participants were business and economics students. The remaining participants had a background in psychology, law, arts and culture, health sciences, European studies, or philosophy. Demographics such as age, field of study, and nationality do not differ in the four treatments.

the target of a stereotype threat in the second stage. Note that this is the case for men in both treatments with a stereotype threat against women (IST-W and EST-W) and for women in the treatment with a stereotype threat against men (EST-M).

Hypothesis 2. *In IST-W and EST-W, for men there is a significant increase in performance from the first to second stage. For women this is the case in EST-M.*

Stereotype threat theory states that the performance of individuals belonging to a stereotyped target group is harmed relatively to a situation entailing no stereotype threat (e.g., Steele & Aronson, 1995). Therefore, the adverse effect of a stereotype threat should counteract the expected positive effect of the competitive incentives on performance for women in the IST-W and EST-W treatments, and for men in the EST-M treatment. However, before we can state our hypotheses regarding men and women in these treatments, we need to (i) verify that men and women from our participant pool do indeed hold the stereotype that men are better at mathematics, and (ii) measure if and how men and women from our participant pool update their belief about men's and women's mathematics performance after having read the quotes we use to explicitly induce stereotype threats.

For this purpose, we conducted a separate incentivized online experiment with respondents who did not participate in the main experiment but are recruited from the same participant pool.[14] In the first part of the online experiment, respondents' receive an explanation of the non-competitive first stage of the main experiment. Given this information, they are asked to estimate the average number of correct answers by respectively male and female participants during that stage in the main experiment. This provides us with information about their potential stereotyped beliefs. The second part of the online experiment is split into two versions. Half of the male and female respondents read the quote that we use in the main experiment to explicitly induce a stereotype threat against women, whereas the remaining half of male and female respondents read the quote that we use in the main experiment to explicitly induce a stereotype threat against men. Subsequently, in both versions, respondents are asked whether they want to change their initial estimates. This provides us with information about the effectiveness of our explicit stereotype induction.[15]

[14]The participants in the online experiment do not differ significantly form the lab participants in terms of age, nationality, and field of study (see the Web Appendix).

[15]Respondents are rewarded for the accuracy of their estimates. See the Web Appendix for the full instructions, which provide the details of the monetary incentives.

Table 2. Online experiment – Average estimate of the gender gap in performance

	Version EST-W (n = 60)		Version EST-M (n = 60)	
	Before info	After EST-W info	Before info	After EST-M info
Male respondents	2.2**	5.8***	1.9**	−2.1***
	(6.4)	(5.2)	(5.8)	(4.5)
Female respondents	2.8**	4.9***	2.7***	−1.4*
	(6.7)	(6.4)	(4.7)	(4.3)

Note: A respondent's estimate of the gender gap in performance equals the respondent's estimate of correct answers of men minus a respondent's estimate of correct answers of women. EST-W (EST-M) refers to the piece of information to explicitly induce a stereotype threat against women (men). The "Before info" part of the questionnaire is the same in both versions. ***, **, and * indicates significant at 1%, 5%, and 10%, respectively, based on Wilcoxon Signed-rank tests for the null hypothesis of no gender gap in performance. Standard deviation in parentheses.

In Table 2, we report male and female respondents' estimate of the gender gap in performance of participants in the main experiment, where the gender gap is defined as a respondent's estimate of correct answers of men minus a respondent's estimate of correct answers of women. Before receiving the stereotyping piece of information (columns 'Before info'), both male and female respondents believe that men perform the multiplication task significantly better than women. That is, these data are consistent with men and women from our participant pool holding the stereotype that men are better at mathematics. This gap is significantly different from zero for both male and female respondents, also after having read the quote that women may indeed be worse in mathematics (column 'After EST-W info'). Thus, we can assume that women in our main experiment face a stereotype threat in the implicit as well as explicit stereotype threat treatment, which leads to our next hypothesis.[16]

Hypothesis 3. *In IST-W and EST-W, the increase in performance of women between stages is significantly smaller than men's increase in performance.*

Moreover, after having seen the quote stating that women are indeed worse than men in mathematics, female respondents' estimated performance gap is even increased relative to the initial stereotyped beliefs.[17] Thus, we conclude that the explicit induction of a stereotype threat against women in EST-W worked as intended and we formulate our next hypothesis.

[16]The hypotheses we present here are also consistent with the model of self-stereotyping proposed by Bordalo, Gennaioli, and Shleifer (2014).

[17]Female respondents: 2.8 vs. 4.9, $p = 0.0017$, Wilcoxon signed-rank test. Our manipulation also worked for men (2.2. vs. 5.8, $p < 0.001$, Wilcoxon signed-rank test), but we do not have an a priori hypothesis regarding men's belief updating.

Hypothesis 4. *For women, the increase in performance between stages is in EST-W significantly smaller than in IST-W.*

Finally, we see in Table 2 that male (and female) respondents reverse their initial belief in a gender gap in performance after having seen the quote stating that men are worse than women in mathematics (column 'After EST-M info'). Thus, in EST-M we explicitly induce a stereotype threat against men as intended.[18] This leads to our hypothesis regarding behavior in the EST-M treatment.

Hypothesis 5. *In EST-M, the increase in performance between stages of men is significantly smaller than the one of women.*

4 Results

In this section, we first report results regarding gender difference in performance in the non-competitive first stage, followed by results related to our hypotheses. Thereafter, we present a robustness check when controlling for attitudes towards risk and competition. Throughout this section, we use either Wilcoxon signed-rank (WSR) tests or Mann-Whitney (MW) tests to investigate differences in distributions. All tests are two-sided.[19]

Figure 1 depicts the cumulative distributions of the number of correct answers during the non-competitive first stage performance pooled across treatments, separately for women and men. The figure shows that the men's performance distribution statistically dominates women's performance distribution, i.e., men perform significantly better than women.[20]

Importantly, the figure also indicates a large inter-individual heterogeneity in performance suggesting that it is important to control for baseline performance, as we do, when investigating the effect of competition on performance in the second stage.

[18] Male respondents (Female respondents): 1.9 vs. -2.1 (2.7 vs. -1.4), $p < 0.001$ in both cases, Wilcoxon signed-rank tests.

[19] We report two-sided tests for all comparisons. For those comparisons where we have a directional hypothesis, the reported tests regarding insignificance are not different when using a one-sided test.

[20] The pooled results for men's and women's average number of correct answers in the first stage are: 28.3 (men) vs. 23.2 (women), $p < 0.001$, MW test, n = 190. We exclude two male participants who participated in treatment EST-W and TRP, respectively, and are clear outliers in terms of performance. These are the only participants who, in both stages, solved all problems available (70) before time elapsed. The results we present are qualitatively the same if we include those two participants.

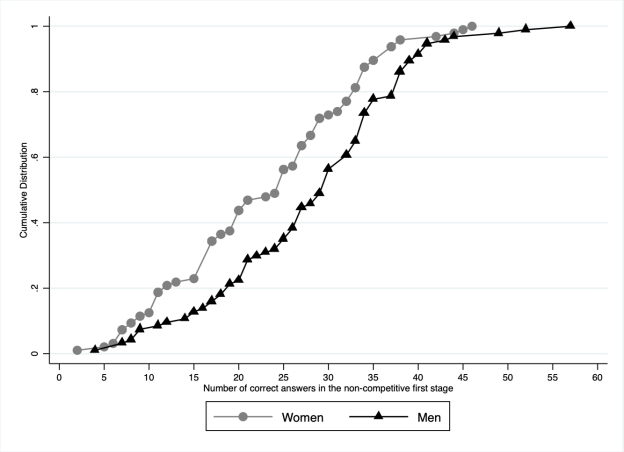


Figure 1. Performance in non-competitive first stage - Women and Men
(Cumulative distributions pooled across all treatments)

4.1 Hypotheses testing

To analyze how men and women respond to competitive incentives in a mixed-gender competition, we make two types of comparisons. First, we compare both men's and women's performance in the non-competitive stage to their performance in the competitive stage. Second, to evaluate possible gender differences in the response to competition, we perform a difference in difference analysis, comparing men's change in performance between stages to women's change in performance.

Table 3 summarizes the performance (i.e, the number of correctly solved multiplications) of women and men in each treatment and also reports the change in performance from the first to the second stage for women and men, within each treatment. In addition, Figure 2 depicts the cumulative distributions in performance change separately for men and women, also in each treatment. In the following, we focus on our main variables of interest which are the change in performance from stage 1 to stage 2 and how this change differs between genders.

We first analyze behavior in the Twice Random Pay (TRP) treatment, which also informs us about whether we have to take into account learning or fatigue effects. In this treatment, men increase their performance by +0.3 correct answers and women decrease it by −0.9 correct answers. None of these differences is significant (women: $p = 0.515$, men: $p = 0.819$, WSR tests).[21] We summarize this observation in our first result.

[21] In addition, we find for both genders a substantial overlap between the cumulative distributions of the first and second stage performance (see the Web Appendix).

Table 3. Performance in non-competitive and competitive stages

	TRP		IST-W		EST-W		EST-M	
	Women	Men	Women	Men	Women	Men	Women	Men
No competition	25.7	30.4	21.8	23.9	23.7	30.8	21.3	27.2
	(11.58)	(10.35)	(10.79)	(11.54)	(9.71)	(10.10)	(8.84)	(9.45)
Competition	24.8	30.7	26.2	28.0	26.7	35.4	23.6	25.6
	(15.33)	(10.13)	(11.25)	(13.23)	(11.12)	(12.97)	(11.38)	(10.33)
Change	-0.9	+0.3	+4.4	+4.1	+3.0	+4.6	+2.3	-1.6
p-value	0.515	0.819	0.019	0.012	0.044	0.004	0.090	0.269

Note: TRP ... Twice Random Pay; IST-W ... Implicit Stereotype Against Women; EST-W ... Explicit Stereotype Against Women; EST-M ... Explicit Stereotype Against Men. Numbers indicate average number of correct answers, standard deviations in parentheses. Change equals performance in competitive second stage minus non-competitive first stage, p-values are based on Wilcoxon signed-ranks tests.

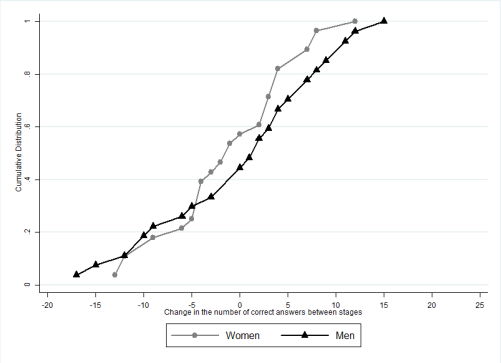
Result 1. *In the Twice Random Pay treatment, men's and women's change in performance between stages is not significant.*

This result supports our first hypothesis and shows that there is neither a fatigue nor an experience effect for both genders. For completeness, we also note that the change in performance does not differ between women and men ($p = 0.582$, MW test; see also Figure 2a). Therefore, it is safe to assume that repetition does not effect the performance of our participants.

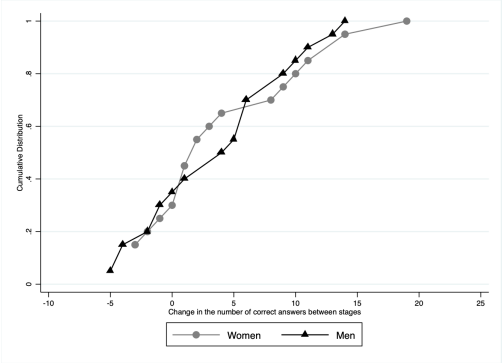
Hypothesis 2 states that when there is no stereotype threat then both men and women increase their performance from the non-competitive to the competitive stage. Men do not face a stereotype threat in the treatments IST-W and EST-W and Table 3 shows that in both treatments men's performance indeed increases significantly. For women, there is no stereotype threat in EST-M and their performance is increased in this treatment, albeit only at the 9% significance level. Thus, our second hypothesis is largely supported and we summarize in our next result.

Result 2. *When there is no stereotype threat, both men and women increase their performance (marginally) significantly from the non-competitive to the competitive stage.*

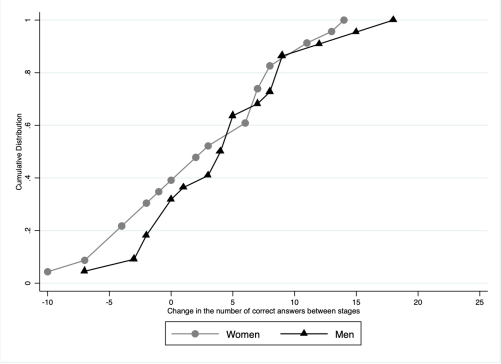
Next we analyze if Hypothesis 3 is supported which refers to a comparison of the change in performance of men and women between stages. Specifically, it states that whenever there is an implicit or explicit stereotype threat against women, then their increase in performance (if any) is smaller than the increase in performance of men. Figures 2b and 2c depict the cumulative distributions of performance changes of men and women in both treatments with a stereotype threat against women (IST-W and EST-W). For both treatments, the distributions show a significant overlap, suggesting that there is no difference in performance change between men and women. This impression is corroborated by statistical



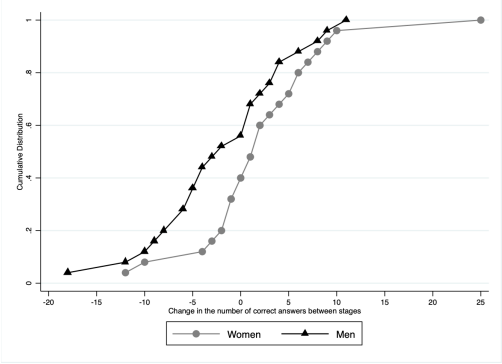
(a) Twice Random Pay (TRP)



(b) Implicit Stereotype against Women (IST-W)



(c) Explicit Stereotype against Women (EST-W)



(d) Explicit Stereotype against Men (EST-M)

Figure 2. Performance change between first and second stage – Women and Men (Cumulative distributions separate by treatments)

tests. In both treatments the changes in performance of men and women are indistinguishable (IST-W, $p = 0.968$; EST-W, $p = 0.510$; MW tests). Thus our third hypothesis is not supported and we summarize this finding in the following result.

Result 3. *When there is a stereotype threat against women, there is no difference between men an women in the change of performance between stages.*

Importantly, the reason that there is no difference between men and women is that the latter, just as men, significantly increase their performance from the non-competitive to the competitive stage. This can be seen from Table 3 which reports a significant performance increase of women in both treatments with a stereotype threat against women (IST-W: $p = 0.019$, EST-W: $p = 0.044$; WSR tests). We summarize these observations in our next result.

Result 4. *Women significantly increase their performance between stages even if there is stereotype threat against women.*

Note that the last two results are in contrast to a number of studies showing that women do not increase their performance in response to competitive incentives or do so less strongly than me do.

The final hypotheses regarding women behavior compares their change in performance from the non-competitive to the competitive stage when faced with an implicit versus an explicit stereotype threat (Hypothesis 4). Figure 3 shows the cumulative distributions of women's performance changes between stages for both treatments. There is a considerable overlap of the distributions suggesting that there is no difference, which is corroborated by a statistical test ($p = 0.600$, MW test). Thus, we do not find support for Hypothesis 4.

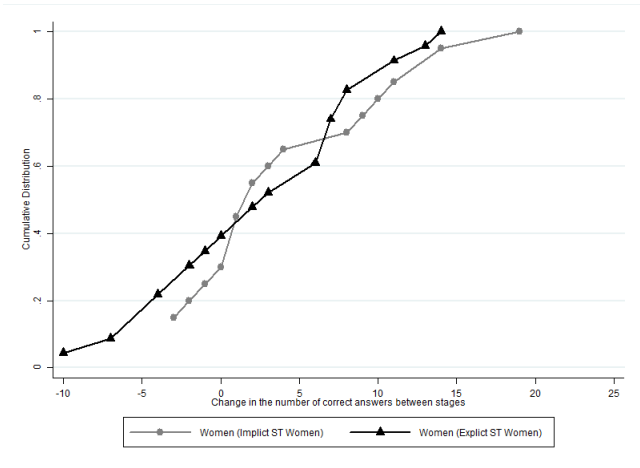


Figure 3. Change in performance from non-competitive to competitive stage - Women only (Cumulative distributions for IST-W and EST-W)

Result 5. *Women's change in performance between stages is not significantly different in the Implicit and Explicit Stereotype Threat treatments.*

Our last hypothesis compares the performance change of men and women between stages in the treatment with an explicit stereotype threat against men. Figure 2d shows the cumulative distributions of these changes for men and women. We see that men's distribution consistently lies above women's distribution pointing at a difference between genders. A MW test returns that the difference is significant at $p = 0.066$. Thus our Hypothesis 5 receives marginal statistical support. Interestingly, the difference in performance change when there is an explicit stereotype threat against men has two sources. First, in contrast

to all other competition treatments, men do not increase their performance from the non-competitive to the competitive stage; in fact, their performance at the competitive stage is insignificantly worse than at the non-competitive stage (see Table 3, EST-M: Men). Second, women marginally significantly increase their performance between stages (see Table 3, EST-M: Women).

Result 6. *When there is an explicit stereotype threat against men, women increase their performance marginally significantly more than men do. This holds because men do not increase their performance, whereas women do increase it.*

In a nutshell, we find supportive evidence for Hypotheses 1 and 2 and, at the marginal significance level, also for Hypotheses 5. In contrast, Hypotheses 3 and 4 are rejected. The reason why we do not find support for the latter two hypotheses is because women respond significantly positive to competitive incentives despite an implicit or even explicit stereotype threat against them. This is in contrast to what could be expected given the published literature on how women respond to competitive incentives.

We next discuss if the presented results are robust when controlling for some individual characteristics that are arguably important in environments with competitive incentives. An important feature of the competitive stage is that the payment participants receive depends not only on their own performance but also on the performance of another participant. Therefore, gender differences in confidence levels regarding their relative performance might affect behavior under competition. Moreover, possible gender differences in risk attitude (e.g., Croson & Gneezy, 2009) and competitive attitude might also cause gender differences in performance, especially at the competitive stage.

Recall that we elicited confidence levels by asking participants to rank their performance in the first stage by indicating the likelihood that 0, 1, 2, 3, or 4 of four other randomly chosen participants performed better at this stage. We find a clear gender difference in confidence levels. Specifically, 19.6% of men believe that none of the other four participants were performing better, whereas for women this the case only for 8.7%. Equivalently, 21.3% of women believe that they performed worse than all four members of the comparison group, but for men this number amounts to only 7.7%. These gender difference in confidence levels is statistically significant. Moreover, using the risk questionnaire from Dohmen et al. (2011) and a competition attitude measure based on the Mach-IV test, we also find that women are significantly less risk tolerant and less competitive than men.[22]

[22]For a detailed discussion of the results regarding confidence, risk attitude, and competitive attitude, see the Web Appendix.

To determine if and how these factors jointly affect the change in performance between stages, and evaluate whether the treatment effects on men's and women's competitive performance are robust to controlling for these factors, we run regression analyses, which are reported in Table 4. In specifications (1) and (2) the Twice Random Pay treatment is the omitted category, in specification (3) the omitted category are men in the Twice Random Pay treatment, and in specification (4), which looks at women's performance only, it is the IST-W treatment. The variables *IST-W*, *EST-W*, and *EST-M* are dummy variables representing the three competition treatments. *Women* is a dummy variable taking on value 1 when the participant is female and the variables *Women* × *IST-W (EST-W) [EST-M]* interact female participants with each of the competition treatments. *Per-non-comp* is a continuous variable that controls for performance at the non-competitive first stage and *Confidence*, *Risk*, and *Competitive* are continuous control variables for the respective individual characteristics discussed above.[23]

Specifications (1) and (2) separately investigate the behavior of men and women, respectively. The coefficient estimates show that the (in)significant results regarding men's and women's behavior that we identified above are robust. The same holds when we directly compare men's to women's response to competition—see the (in)significant interaction terms in specification (3)—and when we directly compare women's competitive response in the EST-W treatment to their response in the IST-W treatment (see the insignificant coefficient of *EST-W* in regression (4)). These observations lead to our next result.

Result 7. *The treatment effects regarding the change in performance between stages are robust when controlling for confidence level, risk attitude, and competitive attitude.*

4.2 Is there a stereotype threat shadow? A robustness check

Overall, the results we obtained are not in line with stereotype threat theory, except for the treatment with an explicit stereotype threat against men. As a robustness check, we look at the data from a different angle, namely by contrasting gaps in performance between groups. The idea is to compare the performance of two groups engaging in the same task in different situations. In one situation, no group faces a stereotype threat and, in another situation, one group is negatively stereotyped whereas the other group is not. Stereotype

[23] *Confidence* can take on values 0, 1, 2, 3, and 4 with lower numbers indicating higher confidence, *Risk* can take on integer values from 0 to 10 with higher numbers indicating higher risk tolerance, and *Competitive* can take on integer values ranging from 20 to 140 with higher scores indicating stronger competitive attitudes.

Table 4. Linear regression model of change in performance between stages

Dependent variable:	*change in performance between stages*			
	Men	Women	Pooled	Women
	(1)	(2)	(3)	(4)
IST-W	2.930*	5.462***	3.668**	base group
	[1.887]	[2.064]	[1.857]	
EST-W	4.302**	4.051**	4.461**	-0.780
	[1.885]	[2.067]	[1.861]	[2.001]
EST-M	-2.008	3.174*	-1.496	n.a.
	[2.162]	[1.825]	[2.051]	
Women			-0.553	n.a.
			[1.877]	
Women × IST-W			1.959	
			[2.699]	
Woman × EST-W			-0.085	
			[2.693]	
Woman × EST-M			4.964*	
			[2.706]	
Perf-non-comp	-0.064	0.107	0.263	-0.020
	[0.088]	[0.061]	[0.051]	[0.113]
Confidence	-0.550	1.476*	0.662	1.539
	[1.102]	[0.757]	[0.610]	[1.159]
Risk	0.749**	0.158	0.470*	0.704
	[0.301]	[0.376]	[0.239]	[0.471]
Competitive	0.089*	0.103	0.093**	0.010
	[0.050]	[0.073]	[0.046]	[0.099]
Intercept	3.29	0.400	1.084	7.161
	[4.627]	[8.008]	[4.293]	[10.551]
Other controls	YES	YES	YES	YES
Observations	94	96	190	43
R^2	0.246	0.157	0.173	0.130

Note: The continuous variables are demeaned for better interpretation of the coefficient values (e.g., *Non-competitive performance*$_i$ = actual non-competitive performance$_i$ - sample mean of non-competitive performance). Other controls include age, nationality, and field of study. ***,** , and * significant at 1%, 5%, and 10%, respectively. Robust standard errors in brackets.

threat theory would predict that under such circumstances the difference in performance between the two groups differs between the two situations because the performance of the stereotyped group is harmed in the situation entailing a stereotype threat.

In the non-competitive stage, at which neither men nor women are exposed to a stereotype threat, we observe that the gender gap in performance favors men in each treatment. Stereotype threat theory would predict that the gender gap in performance (i) increases in the second stage of the IST-W and EST-W treatments, (ii) decreases in the second stage of the EST-M treatment, and (iii) stays the same at the second stage of the TRP treatment. To test these predictions, we compare in each treatment the gender gap in performance in the second stage to the gender gap in performance in the first stage.

Figure 4 displays the average gender gap in performance at each stage, separately for each treatment.[24] The figure clearly shows that in the TRP treatment, with no stereotype threat at both stages, and in both treatments with a stereotype threat against women at the second stage (IST-W and EST-W), the difference of the gender gaps between stages are small and insignificant (TRP: $p = 0.580$, IST-W: $p = 0.970$, EST-W: $p = 0.603$; WSR tests).[25]

In contrast, in the EST-M treatment, with an explicit stereotype threat against men, the average gap in performance between genders is substantially reduced in the second stage, and the reduction is significant at $p = 0.082$ (first stage vs. second stage: 6 vs. 2). We summarize these observations in our next result.

Result 8. *When there is a stereotype threat against women, the observed change in the gender performance gap between the non-competitive and competitive stage is not consistent with explanations based on stereotype threat theory, but it is consistent with it when there is an explicit stereotype threat against men.*

Overall, our results do not support the common finding that women's performance is impaired in situations entailing a mathematical stereotype threat against women. However, we do find evidence for the adverse effect of a stereotype threat on performance of men when we explicitly induce a stereotype threat against men. The relative worsening of men's performance in our setting when informed about women's superior ability in mathe-

[24]The average gender gap in performance at the second stage equals $\sum_{i=1}^{n}(2^{nd}\textit{stage performance male}_i - 2^{nd}\textit{stage performance female}_i)\frac{1}{n}$, in which n is the total number of pairs that are formed for the competition in the second stage. The average gender gap in performance at the first stage equals $\sum_{i=1}^{n}(1^{st}\textit{stage performance male}_i - 1^{st}\textit{stage performance female}_i)\frac{1}{n}$, using the pairs that are formed for the second stage. To compute the gender gaps in the TRP treatment, we use in both stages the pairs that we form to determine who gets paid in the second stage.

[25]The average gender gaps at stage 1 and stage 2 are: TRP: 4.9 vs. 6, IST-W: 2.1 vs. 1.8, EST-W: 7.5 vs. 8.5.

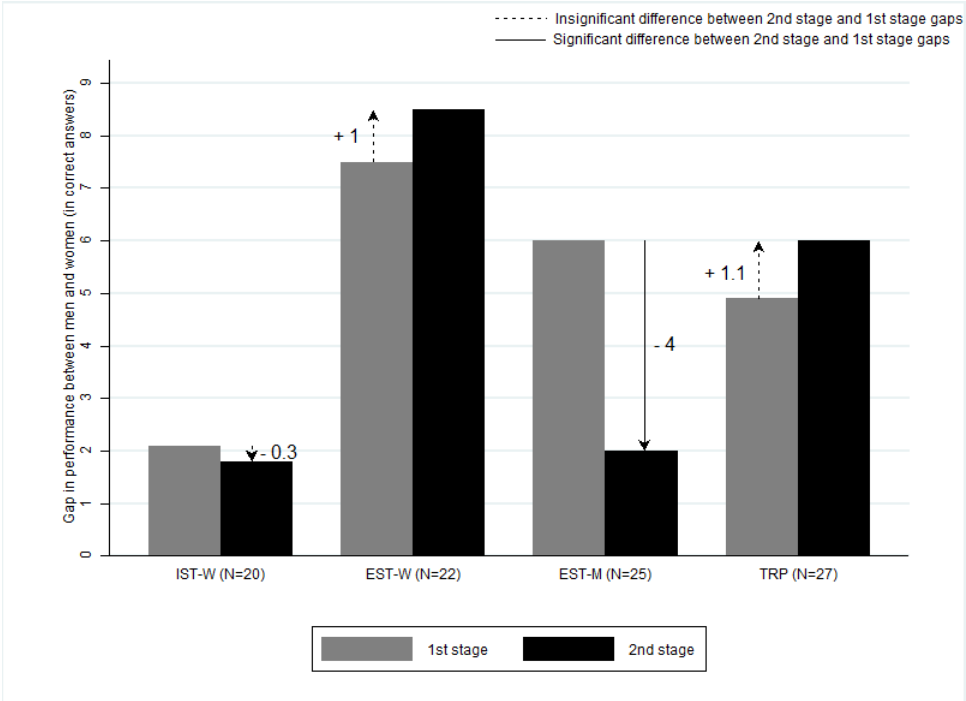


Figure 4. Gender gap in performance at each stage

matics is significant, even though men are not considered to be a group that is prone to a mathematical stereotype threat.[26]

4.3 Attempts and accuracy

Changes in performance between stages may be due to two non-exclusive reasons. It could be that participants make fewer attempts to solve the multiplication problems or that they make more errors, that is, that their accuracy rate of attempts decreases. In this section, we provide an exploratory analysis of these possibilities.

Table 5 reports participants' average number of attempts to solve problems and their accuracy rate, defined as the number of correct answers over the number of attempts, in both stages. For both measures, the table also shows the difference between stages and whether these differences are significant. In the TRP treatment, as one would expect, neither men's nor women's number of attempts and accuracy rate differ significantly between

[26]According to stereotype threat theory, the worsening of performance should be especially observed if the targeted group is prone to the stereotype threat. However, proneness is not a necessary condition to experience stereotype threat. For instance, Aronson et al. (1999) find that Caucasian men perform worse in math-tests when informed about male Asian-Americans' superior ability in mathematics.

Table 5. Attempts and accuracy rate

		Average number of attempts			Average accuracy rate		
		1st stage	2nd stage	Difference	1st stage	2nd stage	Difference
TRP	Women	31.35	31.65	0.30	0.851	0.821	-0.030
	Men	36.44	37.07	0.63	0.817	0.822	0.004
IST-W	Women	28.55	33.05	4.5^{***}	0.737	0.779	0.042
	Men	29.4	32.95	3.55^{***}	0.802	0.841	0.039
EST-W	Women	29.35	32.13	2.78^{***}	0.795	0.823	0.028
	Men	35.86	41.55	5.69^{***}	0.848	0.836	-0.012
EST-M	Women	26.76	29.64	2.88^{***}	0.752	0.758	0.006
	Men	31.72	31.80	0.08	0.850	0.785	-0.065^{***}

Note: A participant's accuracy rate equals the number of correct answers divided by the number of attempts to solve problems. We do not include two female participants in the Twice Random Pay treatment because they did not attempt any question in the second stage (we interpret their behavior as indicative of boredom). *** and ** significant at 1% and 5%, respectively, derived from Wilcoxon signed-rank tests.

stages. In contrast, in both treatments with a competitive second stage and a stereotype threat against women (IST-W and EST-W) both men and women significantly increase their number of attempts without jeopardizing their accuracy of answers, as the latter does not change significantly. Interestingly, the pattern is very different in the treatment where we induce a stereotype threat against men (EST-M). There, women significantly increase their number of attempts from the first to the second stage, whereas men do not. In addition, the accuracy of answers is unchanged for women but decreases significantly for men.

Result 9. *When there is competition in the second stage, women increase their number of attempts without a change in the accuracy rate, irrespective of the stereotype threat. In contrast, when there is a stereotype threat against men, men do not increase their number of attempts and exhibit a decrease in the accuracy of answers.*

The distinct behavior of men in the EST-M treatment invites further scrutiny. In what follows, we shed some more light on the performance of men in this treatment. Specifically, we examine men's change in performance between stages broken down into three periods of 100 seconds to test whether the poorer performance of men persists over the whole 5 minutes of competition.

Table 6 reports men's change in the number of attempts and accuracy rates between stages for those three periods, separately for the treatments where they face no stereotype threat (IST-W & EST-W) and for the treatment where they do face one (EST-M). The table

also shows the difference in changes between treatments and whether these differences are significant. We observe that in IST-W & EST-W the number of attempts to solve problems increases between stages in all three periods, whereas in EST-M it decreases in the first two periods and is weaker in the third period. A comparison of changes between treatments show that these differences are statistically significant in the first two periods. Similarly, in IST-W & EST-W the accuracy rate is increasing between stages in all three periods, whereas in EST-M it decreases over the whole duration of the performance. Between treatments the changes are statistically significant in the first period. We summarize these observations in our last result.

Table 6. Men's change in attempts and accuracy between stages

	Change in number of attempts between stages			Change in accuracy rate between stages		
	0–100	101–200	201-300	0–100	101–200	201–300
EST-M	-0.88	-0.56	1.52	-0.148	-0.036	-0.040
IST-W & EST-W	0.64	1.55	2.48	0.027	0.028	0.029
Difference	-1.52^{**}	-2.11^{***}	-0.96	-0.175^{***}	-0.064	-0.069

Note: 0–100, 101-200, 201-300, respectively, refer to the initial, middle, and final 100 seconds of the 5 minutes performance. We pool IST-W and EST-W because men's change in the number of attempts between stages is not significantly different across periods in these treatments. The same holds for the change in the accuracy rate between stages. *** and ** significant at 1% and 5% level, respectively, based on Mann-Whitney tests.

Result 10. *In comparison to situations without a stereotype threat against men, when there is an explicit stereotype threat, men's performance in solving problems is lower because they tend to make fewer attempts and have a decreased accuracy rate. These differences are especially pronounced at the beginning of the task.*

5 Discussion and conclusions

In this paper, we assess the effect of stereotype threats on the competitive performance of men and women. To our knowledge, this is the first experimental study in economics to investigate not only stereotype threats against women but also against men. Specifically, we conduct a controlled laboratory experiment to test the hypothesis that stereotype threats against men and women, respectively, harm their performance in mixed-gender competition.

For women, our results do not corroborate the psychology research on stereotype threat. According to this literature, we should expect an increased gender gap in performance in situations where the threat exists compared to a situation where this is not the case. In

contrast, in our experiment women's performance under competition appears not to be hampered, neither by an implicit nor an explicit stereotype threat. Interestingly, however, when inducing an explicit stereotype threat against men, we do see an adverse effect on the performance of men.

Our main interest was to examine if stereotype threat impedes performance under competitive incentives. The latter element may explain why our results for women differ from the psychology literature because it studies mathematical stereotype threats against women exclusively in non-competitive situations. Consistent with this interpretation, we do observe that women perform worse than men in our non-competitive task, when assuming an implicit stereotype threat against women at this stage. However, we are doubting that this interpretation would survive scrutiny because it appears unlikely that a gender-based stereotype threat has an effect in a non-competitive environment, but does not have an effect in a competitive environment where it is known to women that they are competing with men. Another possible reason why we do not observe an adverse effect of stereotype threat on women's performance in our study is that we use monetary incentives. This element might wipe out the negative effect of stereotype threat identified in psychology research.[27]

We do not find that women respond less strongly than men to competitive incentives, which contrasts existing results in the economics literature indicating that men respond more strongly to competitive incentives than women do (e.g., Gneezy et al., 2003). In our three competitive environments, we observe in two treatments that men and women react similarly to competitive incentives, and in a third treatment women even react stronger than men do. One possible explanation for this diverging result relates to the observed large inter-individual heterogeneity in (baseline) performance. Most other studies use between-subjects designs which make it impossible to control for performance heterogeneity at the individual level. Our within-subjects design takes care of this and it may be one reason why we do note see a gender difference in response to competition.

Another possible explanation for the contrasting results is that we control to a larger extent for the risk involved in competition than other studies do. In previous studies, competitive incentives are commonly compared to risk-free piece-rate incentives. However, tournament incentives clearly also include a risk component. Thus, it is likely that part of the gender difference in response to competition is due to the well-documented difference in risk attitudes between men and women (e.g., Croson & Gneezy, 2009).[28] To control for this,

[27]Using monetary incentive in a non-competitive situation, Fryer, Levitt, and List (2008) did not find evidence for the negative impact of stereotype threat either.

[28]See also, Gillen, Snowberg, and Yariv (2019) who argue that the gender difference in tournament entry

we included a risk-component in our non-competitive treatment by paying out the achieved outcome with only a 50% chance. If risk-attitudes are a major component for performance under competition, this design feature may explain why we do not find a difference between men and women under competition.

An interesting and novel result is that when men are facing a stereotype threat against them, they decrease both the number of attempts to solve problems as well as the accuracy rate in solving problems they do try. The former effect is consistent with that a stereotype threat reduces the effort men are willing to put into the task. The lower accuracy rate, however, cannot be explained with lower effort. A possible explanation for it is 'distracting thoughts', which have been shown to interfere with working memory and attention that are fundamental for good performance of the multiplication task (e.g., Brewin & Smart, 2005). Accordingly, the cause for men's lower accuracy level under competitive incentives when facing a stereotype threat is that they are distracted by the information they receive just before the start of the competition that contradicts a stereotype they hold—namely, that they are worse and not better than women at math.

This distracting thoughts explanation is consistent with the results in the other treatments. They are unlikely to affect men's performance in the implicit and explicit stereotype treatment against women treatments or women's performance in the explicit stereotype against men treatment because in these treatments men and women, respectively, are not the target of the stereotype threat. Women in the implicit and explicit stereotype against women treatments are unlikely to be affected by distracting thoughts because the text they read does not contradict the prior belief women hold about the invoked stereotype.

One might wonder whether an experimenter demand effect (Zizzo, 2010) has driven the results regarding men's underperformance when we induce a stereotype threat against them. While we cannot fully exclude this possibility, we believe that it is unlikely to be a main driving force. If the information we provide induced a strong experimenter demand effect leading to worse performance, we would expect that this holds for men as well as for women. However, in the explicit stereotype against women treatment there is no indication for such an effect.

The findings of our study can have practical implications for policy design that copes with stereotype threat, for instance at the workplace. A recommendation found in the stereotype threat literature to prevent adverse effects of stereotype threat is the "stereotype nullification", i.e., to explicitly provide individuals with information that does not conform to the

identified by Niederle and Vesterlund (2007) vanishes, when properly controlling for risk preferences.

stereotype they hold (e.g., Smith & White, 2002). However, our results indicate that, at least within a competitive context, no information about the stereotype should be provided at all. If men and women are already competing against each other, they appear to cope well with an eventual stereotype they hold in terms of performance. Our results suggest that providing information contradicting a stereotype seems to be irrelevant for the performance of the a priori negatively stereotyped group, but could harm the performance of the a priori positively stereotyped group.

Our paper contributes to the research aiming at understanding why women are underrepresented in many high-status jobs and earn lower wages than men. When considering the pervasiveness of stereotype threat, our results indicate that women do not have a weaker reaction to competitive pressure regarding performance when they have to compete against men. Therefore, in contrast to previous conclusions in the literature, women's weaker attitude towards competition does not seem to be an explanation for observed gender differences at the workplace if we consider the context in which men and women are already competing. It would be an interesting avenue of future research to investigate if our conclusions also hold for situations where men and women have to decide whether they should compete or not (Niederle & Vesterlund, 2007).

References

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*(1), 29–46.

Babcock, L., & Laschever, S. (2003). *Women don't ask*. Princeton: Princeton University Press.

Black, S. E., & Strahan, P. E. (2001). The division of spoils: rent-sharing and discrimination in a regulated industry. *American Economic Review*, *91*(4), 814–831.

Blau, F. D., Ferber, M. A., & Winkler, A. E. (2013). *The economics of women, men and work*. Pearson Higher Ed.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2014). *Stereotypes* (Working Paper No. 20106). National Bureau of Economic Research.

Brewin, C. R., & Smart, L. (2005). Working memory capacity and suppression of intrusive thoughts. *Journal of Behavior Therapy and Experimental Psychiatry*, *36*(1), 61–68.

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

Christie, R., & Geis, F. L. (2013). *Studies in Machiavellianism*. London: Academic Press.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474.

Directorate–General for Research and Innovation (European Commission). (2019). *She figures 2018*. Online Publication. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/9540ffa1-4478-11e9-a8ed-01aa75ed71a1

Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review*, *101*(2), 556–590.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550.

Dreber, A., Von Essen, E., & Ranehill, E. (2011). Outrunning the gender gap—boys and girls compete equally. *Experimental Economics*, *14*(4), 567–582.

Eurostat. (2020). *Gender pay gap statistics*. Online Publication. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Gender_pay_gap_statistics

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Ex-*

perimental Economics, *10*(2), 171–178.

Fryer, R. G., Levitt, S. D., & List, J. A. (2008). Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *American Economic Review*, *98*(2), 370–375.

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

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, *94*(2), 377–381.

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

Günther, C., Ekinci, N. A., Schwieren, C., & Strobel, M. (2010). Women can't jump?—An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, *75*(3), 395–401.

Hernandez-Arenaz, I. (2020). Stereotypes and tournament self-selection: A theoretical and experimental approach. *European Economic Review*, *126*, 103448.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139–155.

Inzlicht, M., & Schmader, T. (2012). *Stereotype threat: Theory, process, and application*. Oxford: Oxford University Press.

Iriberri, N., & Rey-Biel, P. (2017). Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision. *Journal of Economic Behavior & Organization*, *135*, 99–111.

Iriberri, N., & Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal*, *129*(620), 1863—1893.

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

Offerman, T. (1997). *Beliefs and decision rules in public good games*. Dordrecht: Springer Science+Business.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*(3), 440–

452.

Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, *10*(5), 1189–1213.

Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex roles*, *47*(3-4), 179–191.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*(1), 415–437.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811.

van Dolder, D., van den Assem, M., & Buser, T. (2020). Gender and willingness to compete for high stakes. *Working Paper, available at SSRN: https://ssrn.com/abstract=3537678 or http://dx.doi.org/10.2139/ssrn.3537678*.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98.