Große, Niels Daniel; Riener, Gerhard

Working Paper

Explaining gender differences in competitiveness: Gender-task stereotypes

Jena Economic Research Papers, No. 2010,017

Provided in Cooperation with:
Max Planck Institute of Economics

Suggested Citation: Große, Niels Daniel; Riener, Gerhard (2010) : Explaining gender differences in competitiveness: Gender-task stereotypes, Jena Economic Research Papers, No. 2010,017, Friedrich Schiller University Jena and Max Planck Institute of Economics, Jena

This Version is available at:
https://hdl.handle.net/10419/32599

JENA ECONOMIC RESEARCH PAPERS

2010 – 017

Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes

by

Niels D. Grosse
Gerhard Riener

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University and the Max Planck Institute of Economics, Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Max Planck Institute of Economics
Kahlaische Str. 10
D-07745 Jena
www.econ.mpg.de

Explaining Gender Differences in Competitiveness: Gender-Task Stereotypes

Niels D. Grosse*and Gerhard Riener†

Friedrich-Schiller-University Jena

February 2010

Abstract

Gender-specific patterns of self-selection into competitive and cooperative environments may have multiple reasons. One of the most prominent explanations to this point is, that there are inherent differences between men and women when it comes to preferences regarding competition. We take a different point of view and claim that gender-task stereotypes are able to explain a large part of the under-representation of women in tournament like environments. We conduct an experiment with a quantitative task which has been shown to have a strong male connotation and a verbal task which we hypothesize to be gender neutral. After controlling for differences in performance, risk attitudes, and overconfidence, we find that women self-select significantly less into competition against men only in the quantitative task. This finding suggests that remaining gender differences for entry into competition are driven by gender-task stereotypes. As a robustness check, we explore the self-selection into incentive schemes given different gender compositions of groups and self-selection into single-sex groups given different incentive schemes. Furthermore, we report the results of a framed field experiment, where we explore a further task – throwing balls into a bucket – that has as well a male connotation. These additional results further strengthen our interpretation.

Keywords: Competition, piece rate, revenue sharing, gender-task stereotype, experiment

JEL Code: C91, J16, J24, M52, D81

*University of Jena, Graduate College "The Economics of Innovative Change," Carl-Zeiss-Str. 3, D-07743 Jena (Germany), Phone: +49 3641 94270, Email: niels.daniel.grosse@uni-jena.de

†University of Jena, Graduate College "The Economics of Innovative Change," Carl-Zeiss-Str. 3, D-07743 Jena (Germany), Email: gerhard.riener@uni-jena.de

We are indebted to Stefan Bauernschuster, Katharina Eckartz, Werner Güth, Marcela Ibanez D., Oliver Kirchkamp, Patrick Nolen, Vera Popova and David Reinstein for insightful comments that helped to improve this paper. Thanks also to Simon Wiederhold, Bastian Rake and Alexander Schacht for enabling the field experiment. Djamila Koberstein, Jonathan Leupert, Conny Malz and Sabrina Vieth provided excellent research assistance.

1 Introduction

> "So my best guess, to provoke you, of what's behind [women's underrepresentation in the science and engineering workforce] is that the largest phenomenon, by far, is the general clash between people's legitimate family desires and employers' current desire for high power and high intensity, that in the special case of science and engineering, there are issues of intrinsic aptitude, and particularly af variability of aptitude, and that those considerations are reinforced by what are in fact lesser factors involving socialization and continuing discrimination."
>
> Lawrence H. Summers, former President of Harvard University, 14 January 2005
>
> (cited from Steele et al., 2007)

The segmentation of labor markets with respect to gender is widespread. Recent experimental studies ascribe part of this segmentation to different self-selection behavior of men and women into competitive environments (Niederle and Vesterlund, 2007; Booth, 2009, for an overview) or performance differences in competitive environments (Gneezy et al., 2003; Gneezy and Rustichini, 2004). Explaining this residual finding by differences in preferences proves to be rather unsatisfactory for guiding policies that try to change the observed behavior. In this paper we therefore try to uncover underlying reasons for theses observed differences by disentangling behavior in competitive environments *per se* (e.g., Gneezy et al., 2003) from purely task-related factors such as *gender stereotypes* or gender-job associations (Akerlof and Kranton, 2000).

Akerlof and Kranton (2000) point out that women might shy away from certain jobs because of gender associations of the task. Steele et al. (2007) reviewed the literature and offer task stereotypes as a potential driving force of gender associations. For example, the disproportional underrepresentation of women in quantitative occupations like engineering can well be explained by gender-job associations that are widespread in society.[1] Why women often do not self-promote to leading positions might, moreover, not be related to the supposed unwillingness of women to compete, but to the fact that leadership has a strong male connotation and to a stereotype that men are better leaders (for evidence on prejudices agains female leaders see Eagly and Karau, 1991, 2002). Gender associations might therefore be expressed in actual differences in performance but also in false stereotypes (Phelps, 1972), i.e. wrong or exaggerated perceptions of differences between men and women.

In labor markets, agents can choose among incentive schemes (or career paths) in different jobs. Therefore, field studies face the problem that only subjects who selected themselves into different jobs are observed; a limitation that can be overcome in experiments. For this purpose, we compare performance and self-selection into cooperative and competitive

[1] As the quotation at the beginning of this paper demonstrates.

environments with a *quantitative task* and a *verbal task* – tasks that have different degrees of gender associations. Hence, we randomly assign subjects to incentive schemes and tasks. Additionally, the experimental setup allows us to apply a variety of controls that are hardly measurable in the field: expected earnings, risk attitudes dependent on the environment, and self-assessment of relative performance.[2]

In the quantitative task, our findings support previous results: women shy away from competing in mixed-gender groups. However, this might been an artefact of the applied task. Numerous educational and psychological studies have shown that girls perform worse in general mathematical skills (OECD, 2006).[3] As a consequence, subjects in previous psychological experiments associate mathematics with being a "male" task (Nosek et al., 2002), even if their own performance is above the male average. These gender associations might critically challenge the interpretation of findings in mathematical real effort tasks. The finding that women perform worse in competitive environments and shy away from competition might be due to the perceived stereotype that mathematics is a male task and that men perform better in this task. In constrast, young women across countries consistently perform better in tests of verbal skills (Guiso et al., 2008). Therefore, verbal tasks do not have this strong male connotation and are at least gender-neutral. For the verbal task, we find that sorting decisions into tournaments are no longer significantly correlated with gender. We conjecture that differences in self-selection can be explained to a large extent by task-related factors on top of the stereotype of competitiveness being male (that can be interpreted as gender differences in preferences for competition *per se*). Our findings can be considered as experimental support to model labor market outcomes in terms of identity and gender-job associations as in Akerlof and Kranton (2000).

Our interpretation of the results leads to a different view on optimal policies: If preferences – at least in a traditional economic view – are taken as primal and should not be changed by policy makers, there is only limited scope for policy intervention, and if so it would have to occur at very early stages in life. However, if sorting into competitive environments is driven by wrong perception of groups based on gender-task stereotypes, this opens a field for policy intervention. For example, in a recent experiment Johns et al. (2005) showed that performance differences based on stereotypes can be mediated by providing information about the existence of the stereotype. Whether this holds for the selection into competitive environments has still to be determined.

Our findings show that the applied task matters. It therefore should not be chosen arbitrarily or solemnly on grounds of non-existent differences in performance between sexes. Instead, the perception of a task in a particular society has to be controlled for in gender research. This methodological contribution should help future studies to take

[2] Dohmen and Falk (2010) find that risk preferences, differences in self-perception, personality traits, and social preferences explain sorting into incentive schemes.

[3] 15-year-old girls perform worse in mathematics and consistently report lower self-related beliefs in mathematics in most OECD countries (OECD, 2006). Guiso et al. (2008) find that this gender gap in mathematics is highly correlated with the World Economic Forum's Gender Gap Index.

into account the nature of underlying tasks that consequently allows for a more precise interpretation of empirical findings.

Furthermore, our design allows us to disentangle task-specific learning effects and incentive effects of payment schemes in a between-subject design with a control group that is paid at piece rate in all real effort stages. Gneezy et al. (2003) find gender differences in solving computerized mazes under competitive schemes in a within-subject design. Compared to a piece rate wage men increase their performance significantly in all competitive settings, while women increase their performance only against women. In subsequent studies on gender-specific performance in competitive environments in other studies results are more mixed and depend on task (Gneezy and Rustichini, 2004; Niederle and Vesterlund, 2007), subject pool (Gneezy et al., 2009), and experimental design (Schwieren and Weichselbaumer, 2009). A criticism of within-subject design in real effort tasks. We find that learning effects capture a large part of observed within-subject difference between piece rate and other incentive schemes.

The outline of this paper is as follows: In section 2, the experimental design is presented. Afterwards, we discuss our theoretical framework and derive our conjectures for the experimental analysis. Results of this experiment are reported in section 4. Section 5 presents the results of a field experiment. The final section concludes.

2 Experimental Design

For this experiment, we chose two different tasks: a *summation task (SUM)* – having a strong male connotation (for related research in psychology see Nosek et al., 2002) – and a *word-order task (WO)* – a task that does not evoke gender-specific associations, as confirmed by a pre-test . Each of the tasks is performed under different incentive schemes: *piece rate*, *tournament* and *revenue sharing*. Under *piece rate*, subjects were paid €0.50 for each correct solution independently of other subjects' performance. Tournament and revenue sharing were conducted in groups of four. In the tournament, the best-performing subject earned €1.40 per piece, while the other subjects in the group earned €0.20 per piece.[4] In revenue sharing, subjects' earnings were based on the average of the solutions by all group members and paid at the rate of €0.50 per piece.

In total, we implemented six treatments which combined *incentive* and *task* treatments. In two treatments (*IC-SUM* and *IC-WO*), we analyze how performance under different incentive schemes and self-selection into incentives schemes depend on the prevailing task. For each task we administered a control treatment (*CO-SUM* and *CO-WO*), where piece rate was paid in all stages. This allows to disentangle incentive and learning effects in the real effort task. As an additional check for the importance of gender, we conducted two

[4]There was no tie breaking. If more than one subject reached the highest number of solutions in a group, all subjects with the highest number of solutions received the same high wage per piece.

treatment (*GC-TO-SUM* and *GC-TO-WO*) in which subjects could choose the group composition according to predefined criteria.

The tasks

In the summation task, subjects were asked to add up five two-digit numbers.[5] Participants were not allowed to use a calculator but could use scratch paper and pencil provided by the experimenter. The numbers were randomly drawn and presented to the subjects in the following way:

12	89	77	34	62	___

After subjects had completed one calculation, they were presented with a new one until five minutes had passed. The number of solved calculations was presented on screen. After receiving the instructions, subjects had the possibility to familiarize themselves with the task in a two-minute non-paid trial round.

In the word-order task, subjects were asked to order five words to build a grammatically correct sentence. This task was presented in the following way:

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

Subjects then had to enter the label of the word in the correct order, as shown in the following solution of the previous example:

Word No.	*Word No.*	*Word No.*	*Word No.*	*Word No.*
4	1	3	2	5

After subjects had completed one sentence correctly, they were presented with a new one until four minutes had passed. The time for this task was shortened by one minute in order to keep subjects' earnings in a similar range for both tasks. As in the summation task, subjects had the possibility to become familiar with the task in a two-minute non-paid trial after receiving the instructions.

[5] This task has been used in various studies concerned with competitive behavior (e.g., Niederle and Vesterlund, 2007; Niederle et al., 2008).

Experimental procedure

Overall, the experiment consisted of a pre-experimental questionnaire, a trial round of the task, 10 stages, and a post-experimental questionnaire. Table 1 provides an overview of the experimental procedure. The treatments – presented in columns – are conducted between-subject. The stages – presented in rows – varied within-subject. Stages in which the real effort tasks were carried out were marked with (SUM) and (WO) for the summation task and the word-order task, respectively. The real effort task was repeated six times in each treatment, resulting in 30 minutes of real effort task in the summation task treatments and 24 minutes in the word-order task treatment. In stages 5, 8, 9, and 10, subjects only had to make decisions without performing tasks. During the whole course of the experiment, subjects received no information about the performance of others, their own relative performance, or their payoffs. In all treatments we additionally measured for differences in risk preferences and self-assessment.

The experiment was conducted in spring and winter 2009 at the laboratory of the Friedrich-Schiller-University, Jena. All subjects were undergraduate students of the University from a wide variety of disciplines. Subjects were recruited online via ORSEE (Greiner, 2004). Overall, 416 subjects participated in 26 experimental sessions. All sessions had a balanced gender composition and consisted of 16 subjects[6]. For payment, one of the 10 stages was drawn from a physical urn at the end of the experiment. Additionally, subjects had the opportunity to earn up to €2.50 in bonus questions regarding their self-evaluation. On average, subjects earned €8.36, with a maximum of €43. The outline of the experiment was provided to subjects in printed form. Detailed instructions, the experiment, and questionnaires were computerized using zTree (Fischbacher, 2007). Translated instructions are provided in appendix A.

Stages 1 to 3 (Variation of incentive schemes)

In the first three stages, we investigated the performance of subjects under different incentive schemes with the same task. After subjects had filled out the pre-experimental questionnaire, instructions for the task were presented and the two-minute non-paid trial round was conducted. Detailed instructions regarding the payment regime were provided before each stage on screen. In the treatments *IC-SUM* and *IC-WO*, matching groups of eight (four men, four women) were formed. Within matching groups, subjects were randomly allocated to groups of four. Every incentive scheme (piece rate, revenue sharing, or tournament) was applied in one of the three stages. In the first stage the prevalent payment regime was always piece rate. The order of the *revenue sharing* and *tournament* was randomized over matching groups in stages 2 and 3.

In the control treatments *CO-SUM* and *CO-WO*, piece rate was applied in every stage to control for possible task-specific and gender-specific learning effects in the second and

[6] Except of two where there were 7 male and 9 female due to no-shows.

Table 1: Experimental procedure by between-subject treatments

Treatment	**IC-SUM**	**IC-WO**	**CO-SUM**	**CO-WO**	**GC-TO-SUM**	**GC-TO-WO**
Task	**Summation**	**Word-Order**	**Summation**	**Word-order**	**Summation**	**Word-order**
No. subjects	n=64	n=64	n=64	n=64	n=80	n=80
Stages 1	Piece rate (SUM)	Piece rate (WO)	Piece rate (SUM)	Piece rate (WO)	Piece rate (SUM)	Piece rate (WO)
Stage 2 to 3	*Stratified:* revenue sharing tournament (SUM)	*Stratified:* revenue sharing tournament (WO)	2 x Piece rate (SUM)	2 x Piece rate (WO)	2 x Tournament (SUM)	2 x Tournament (WO)
Stage 4	Scheme choice (SUM)	Scheme choice (WO)	Piece rate (SUM)	Piece rate (WO)	Group choice & tournament (SUM)	Group choice & tournament (WO)
Stage 5	Scheme choice (given random performance)					
Stage 6	Tournament choice (SUM)	Tournament choice (WO)	Tournament choice (SUM)	Tournament choice (WO)	Tournament choice (SUM)	Tournament choice (WO)
Stage 7	Weakest-link (SUM)	Weakest-link (WO)	Piece rate (SUM)	Piece rate (WO)	Weakest-link (SUM)	Weakest-link (WO)
Stage 8 to 10	Gender choice in: *(stratified)* weakest-link, revenue sharing, tournament	Gender choice in: *(stratified)* weakest-link, revenue sharing, tournament	Scheme choice in male group, mixed group, female group	Scheme choice in male group, mixed group, female group	Scheme choice in male group, mixed group, female group	Scheme choice in male group, mixed group, female group

third stage. In the treatments *GC-TO-SUM* and *GC-TO-WO*, *piece rate* was applied in stage 1 and *tournament* in stages 2 and 3 in random groups of four. In stage 3, subjects were compared to others' performance in stage 2. As a consequence, subjects' performance in stage 3 did not have any negative impact on other subjects' payoff. By using this mechanism, we exclude that fairness considerations influenced the performance of subjects in tournaments.

Stage 4 (Incentive scheme choice / only IC-SUM and IC-WO)

In stage 4 of the *IC-SUM* and *IC-WO*-treatments, we examined self-selection into payment regimes. Before performing the task of stage 4, subjects chose whether they wanted to be paid by piece-rate, revenue sharing or tournament. After that, they performed their respective task under the chosen incentive scheme. In cases in which subjects chose revenue sharing or tournament, we had to make sure that a group size of four was always guaranteed. Therefore, subjects' performance in stage 4 was compared to the previous performance of three other members of the matching group under the same incentive scheme and subjects were informed about this. For each subject we also calculated the probability winning in the tournament given subject's previous performance in the tournament stage and 10,000 draws of three reference subjects.

Stage 4 (Group composition choice / only GC-TO-SUM and GC-TO-WO)

Before stage 4 of the treatments *GC-TO-SUM* and *GC-TO-WO*, subjects were asked to select the attributes of their preferred reference group out of a list. For this list, we chose attributes that are well documented to correlate with university students' mating behavior (e.g., Marmaros and Sacerdote, 2006; Mayer and Puller, 2008): age, gender, number of siblings, distance of birth town, sport practice per week, and membership in societies.[7] After choosing one of the attributes, subjects decided which characteristic of the attribute they preferred. There were two distinct ways for the choice of the characteristic: for the attribute gender, subjects chose between male and female. For the other attributes, we used the fact that characteristics can be ordered and therefore have a median. Thus, subjects could determine whether they wanted to play against (or with) subjects above or below the median. This choice was presented to subjects as a choice between halves of the participants, e.g., "the younger half of participants" or "the older half of participants."

After this choice, three suitable participants were chosen for the comparison of subjects' performance. Alternatively, subjects could choose not to select any reference group. In

[7] The criterion *race* was not relevant for our subject pool, as 94% of the students of Jena University are of German origin. We added siblings as this criterion is positively associated with other-regarding preferences (Van Lange et al., 1997). We asked subjects in a survey after the pilot to rate the importance of these criteria. None of the attributes was significantly correlated with performance in the real effort tasks in stage 2 except for number of siblings in the case of the summation task (ρ=0.203, p=0.07). The ordering of the attributes was randomized over subjects.

this case, they were compared to three randomly selected participants. As in the previous stage of these treatments, the performance of stage 2 was used for the reference group. The necessary demographic data for the choice of reference groups was obtained from the pre-experimental questionnaire before subjects knew the procedure of the experiment.[8]

Stages 5 to 7 (Controls)

Previous studies found that women are less willing to take risk in numerous settings and report to be less willing to take risk (Dohmen et al., 2005). Differences in risk taking are supported by a variety of field studies, but are less pronounced in laboratory experiments (Eckel and Grossman, 2008). Gender differences seem to be context-dependent as well (cf. Croson and Gneezy, 2009). Stage 5, therefore, was designed to measure risk preferences of subjects in this particular game: subjects were randomly assigned a number of "solved" solutions between 1 and 20 with equal probability. Then they were asked to choose between a piece rate, tournament, or revenue sharing without knowing the realization of the random number. At the end of the experiment, subjects were additionally asked to rate their willingness to take risk in a questionnaire. This risk-taking question was taken from the GSOEP (Wagner et al., 2007), which has shown to be a valid and reliable measure for general risk attitudes (see Dohmen et al., 2005). We find a significant and positive correlation between the choice of the tournament in stage 5 and the self-reported willingness to take risk in the questionnaire.

In stage 6, subjects had to choose between the following three competitions: In competition 1, €1.40 per solved calculation was paid to the winner, in competition 2 €1.00, and in competition 3 €0.50; the loser pay was €0.20 per solved calculation in all three cases. After this choise, subjects played against all players of the experimental session that chose the same competition simultaneously. Results of this stage are analyzed in a companion paper.

As an additional test of subjects' self-perception and to investigate whether subjects liked to do the task, subjects performed a *weakest-link* payment task in stage 7 in treatments *IC-SUM, IC-WO*, *GC-TO-SUM* and *GC-TO-WO*. For this purpose, subjects were matched randomly in groups of four and then had to perform the task for the last time. Payment was €0.50 per solved calculation of the worst performing group member. In the control treatments, subjects performed their task under piece rate.

Stages 8 to 10 (Only IC-SUM and IC-WO)

In these 3 stages, subjects could chose their preferred reference group for a submission of their performance under revenue sharing, tournament and weakest link, stratified over

[8] We made sure that gender was not a salient characteristic in the pre-experimental questionnaire. The questionnaire asked for year of study, age, gender, number of siblings, birthplace, previous experience in experiments, sports activity and activity in clubs (in this order).

stages 8 to 10. Subjects are asked - given the respective incentive scheme - to choose whether they want to play with men, women or a random group. This serves as a robustness check as we allow for a self-selection into single-sex groups given the incentive scheme and previous performance.

Stages 8 to 10 (Only CO-SUM, CO-WO, GC-TO-SUM and GC-TO-WO)

Stages 8 to 10 were identical in the four treatments *CO-SUM, CO-WO*, *GC-TO-SUM* and *GC-TO-WO*. In each stage subjects chose whether they wanted to be paid their piece-rate performance of stage 1 according to piece rate, tournament, or revenue sharing. In each stage the gender composition of reference group varied. Three subjects of each reference group were drawn to compare their performance in stage 1 with the respective subject's performance in stage 1. In each of those three stages, subjects were informed about the attribute of the reference group. In stage 8 the reference group consisted of male subjects; in stage 9, no attribute was chosen (random group); and in stage 10 the reference group consisted of female subjects only. Every subject, therefore, made a decision about the payment regime in an group where the all partners were either male or female or wher the gender composition was mixed. In these stages we could establish whether subjects choice depends on the reference group and task.

Measure of overconfidence and stereotypes

Self-selection depends on the subject's belief regarding her relative performance. In laboratory environments men tend to be more overconfident.[9] E.g., Camerer and Lovallo (1999) examine the market entry decisions in an experimental setting using real effort tasks (solving trivia quizzes). Their sample comprises only men, and their subjects enter more often than optimal. Dohmen and Falk (2010) specifically look at sorting decisions into different fixed and variable payment schemes. They do not find a high prevalence of overconfidence, but subjects who are overconfident tend to select more often into tournaments. Niederle and Vesterlund (2007) focus on the gender-specific aspects and find that men are substantially more overconfident than women in a mathematical real effort task. Lundeberg et al. (1994) find that gender differences in overconfidence are highly task-dependent.

As an measure of overconfidence and self-perception, we asked subjects to guess their ranks in a group of four. In all treatments, we asked for ranks given subjects' performance in stage 2 and stage 4. In the treatments *IC-SUM* and *IC-WO*, subjects were additionally asked to guess their rank given their stage 1 performance in (a) a group of radomly selected men, (b) a group of randomly selected women and (c) a randomly selected

[9] In non-experimental settings, Barber and Odean (2001) observe male behavior that is consistent with overconfidence in financial markets which leads to more trading and lower net returns in stock market investments compared to women.

mixed-gender group. Subjects received € 1 (€ 0.5 in *IC-SUM* and *IC-WO*) for each correct guess. Thus any gender differences in the relative assessment of the subjects' rank were elicited in an incentivized environment. For the measure of self-assessment in the respective stage, we calculated the subject's optimal rank using 10,000 draws of three subjects and subtracted subject's guessed rank. Positive values of this measure indicate that a subject overestimates her ranking.

3 Conjectures

Choice of payment regime given different tasks

While the interpretation of previous research would suggest that women shy away from competition because of an inherent preference against competitive environments, we suggest to jointly look at the competitive environment and the nature of the task to get at the underlying causes of choice differences. As explained in the previous section, mathematical tasks have strong male connotations , while verbal tasks are more neutral or tend to be favored by women. These associations might feed into subjects' perception about the performance of the different groups in these specific tasks. We therefore postulate the following hypothesis:

Conjecture 1. *Women choose to compete less in the summation task, if they have reason to belief that at least some of their competitors are male. This effect does not exist in the word-order task.*

Choice of payment regimes given different reference groups

If the choice of the payment regime is driven by subjects' perceptions of groups' average performance in each task and the reference groups are fixed, then gender-task associations should be reflected in the choice of the payment regime. Therefore – building on the arguments used for conjecture 1 – we derive the following hypothesis about the choice of payment regimes given different reference groups:[10]

Conjecture 2. *a) In the summation task, women compete less often relative to men in the all-male and the random group, but not in the all-female group. This is driven by gender-task associations in that particular task.*

b) In the word-order task, women ***do not*** *choose to sort less often into competitive payment regimes than men in the all-male, random and all-female group. This means that self-selection is not driven by gender-task associations in that particular task..*

[10] Note that this conjecture is different from a conjecture that would be derived from results taken from the stereotype threat literature (see Steele, 1998; Benjamin et al., 2007, for an application to economic questions) that analyzes the salience of stereotypes and its effect on actual performance. Here we are interested in the assessment of the subjects' relative performance after the task was performed.

An alternative explanation of selection into payment regimes is based on gender differences in competitive behavior *per se* which lead to less self-selection of women into competitive environments in all tasks (Niederle and Vesterlund, 2007). Previous research has observed gender differences only in competition against men (Gneezy et al., 2003; Gneezy and Rustichini, 2004). From this literature we derive an *alternative* hypothesis to conjecture 2:

Conjecture 3. *In the summation task* ***and*** *in the word-order task, women choose to compete less often in the all-male and the random group but not in the all-female group.*

Choice of reference groups given different payment regimes

Given the assumption that women shy away from competition when the task is associated to be male, we can derive conjectures how women will behave when the gender composition of the reference group can be selected.

Conjecture 4. *(a) In the summation task, women will shy away from competition against men and will choose women more often.*

(b) In the word-order task, women will not shy away from competition against men .

4 Results

4.1 Learning and Performance

We start with a description of the performance of subjects under different incentive schemes. First, we test for gender differences in the performance under piece rate. Figure 1 depicts the cumulative distribution function of the performance in stage 1 pooled over all treatments by gender and task. Men performed better in the summation task and, on average, solved 11.17 calculations while women solved 9.84 calculations. This difference in mean performance is significantly different at the 5%-level using a two-sided t-test with unequal variances ($p = 0.039$).[11] The difference in performance is persistent over the whole range of the distribution but is remarkably pronounced in the upper tail. Gender differences in performance stay significant in all real-effort stages of the CO-SUM treatment. Men, on average, perform better than women at the 5% significance level, except for stage stage 6 and 7 where the level of significance drops to 10%.

In stage 1 of the word-order task, men again perform better and solved 15.3 sentences while women solved 13.7 sentences. However, the gender difference in performance is at the verge of conventional significance levels (t-test, $p = 0.097$). The difference in performance are strongest in the lower tail of the distribution (see figure 1) while no

[11] All subsequent t-tests are two-sided and assume unequal variances.

difference is identifiable in the upper tail. Differences in average performance shrink in subsequent stages of the word-order task. In all stages, except for stage 3, we find no gender difference in average performance of the CO-WO treatment at the 10% significance level.

To isolate incentive effects from learning effects, performance in the tournament and revenue sharing schemes are subsequently compared to the control treatments. To account for differences in initial levels of performance, the analysis focuses on absolute differences in performance from stage 1 to stage 2. Mean changes in performance by gender, task and incentive scheme are presented in table 2. Learning effects, i.e. performance changes in the piece rate, are positive and significantly different from zero (t-test, $p = 0.028$ in CO-SUM, $p = 0.001$ in CO-WO). Revenue sharing does not seem to have any significant effect on average performance in both tasks.

In the summation task, men increase their performance significantly under tournament conditions compared to piece rate (t-test, $p = 0.071$) while women do not increase their performance significantly (t-test, $p = 0.62$). Here, only men are responsive to a change in competitive conditions in a mixed-gender group, however, gender differences in the tournament are not significant (t-test, $p = 0.249$). In the word-order task, performance increase in the tournament is not significantly higher than in the control treatment for women *and* men (t-test, $p = 0.636$ for women, $p = 0.666$ for men). In contrast to the summation task, women even raise their performance more under competitive conditions than men. Gender differences in the tournament are again not significant (t-test, $p = 0.279$). While the observations in the summation task support the basic finding of Gneezy et al. (2003) that women shy away from competing against men, the observations in the word-order task yield the insight that gender differences in the tournament depend crucially on the task. Additionally, learning effects in both tasks account for a substantial proportion of observed performance increase while incentive effects seem to be small if identifiable at all.

In the third stage, we compare whether playing against others' past performance does have an effect on subjects' performance in the tournament. For this purpose we focus on the absolute difference in performance from stage 2 to stage 3 in the CO- and GC-treatments. In the GC-TO-WO treatment, we find that women increase their performance significantly compared to the learning effect (t-test, $p = 0.067$) while we do not find any significant changes for men (t-test, $p = 0.976$). In the same treatment, women even outperform men in absolute terms (22.2 vs. 21 sentences), although gender differences are not significant. In the GC-TO-SUM treatment, we do not find any significant incentive effects, but women converge to mens' performance in absolute terms (12.1 vs. 12.4 solutions). If the tournament scheme does not negatively affect co-players, gender differences in performance seem to vanish.

Figure 1: CDF of stage 1 performance by gender

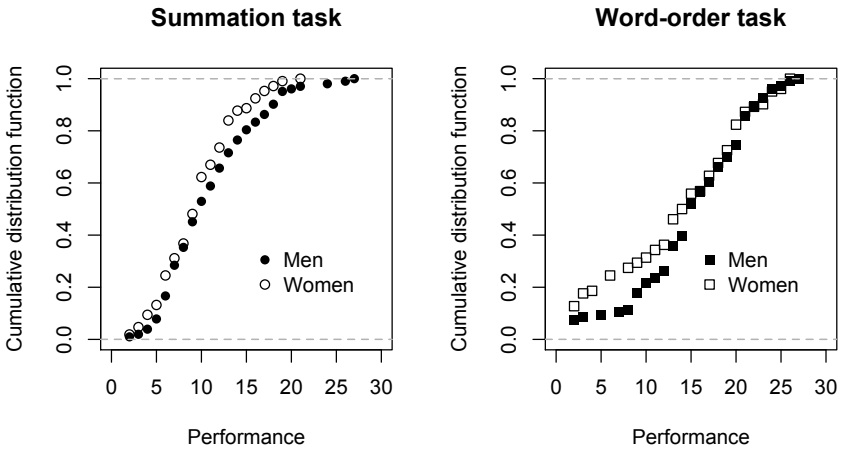

Table 2: Performance change from stage 1 to 2

Task	*Summation task*			*Word-order task*		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
Piece rate	0.66	0.84	0.75	2.00	3.16	2.56
	(0.43)	(0.52)	(0.33)	(0.92)	(1.22)	(0.75)
Tournament	1.67	1.14	1.40	2.54	3.87	3.20
	(0.35)	(0.30)	(0.23)	(0.86)	(0.87)	(0.61)
Revenue sharing	0.73	0.59	0.66	1.75	2.25	2.00
	(0.83)	(0.56)	(0.48)	(1.38)	(1.37)	(0.96)

Note: Mean performance change from stage 1 to stage 2. Standard errors in parenthesis.

Table 3: Choices of payment regimes (IC/Stage 4)

Task	*Summation task*			*Word-order task*		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
Revenue sharing	10.0%	23.5%	17.1%	15.6%	31.3%	23.4%
Tournament	36.7%	11.8%	23.4%	25.0%	15.6%	20.3%
Fisher's exact test	0.046			0.313		

Note: Residual category "piece rate" omitted

4.2 Self-selection into Incentive Schemes

Table 3 depicts the percentage of chosen incentive schemes by gender in stage 4 of the IC-treatments. In both tasks, men select the tournament more often while women tend to prefer revenue sharing. Choices are significantly dependent on gender only in the summation task (Fisher's exact test = 0.046) but not in the word-order task (Fisher's exact test = 0.313). Women tend to shy away from competition only in the mathematical task. However, different choice behavior may also originate from lower performance in the task and therefore lower expected earnings or differences in self-assessment and willingness to take risks.

Therefore, we control for these additional variables in a regression with the choice of tournaments as the dependent variable. Results are presented in table 4. In model (a) dummy variables for tasks (SUM and WO) are interacted with dummy variables for female participants. These interactions replicate the basic findings of the exact tests presented before (p = 0.020 in summation task and p = 0.357 in word-order task). In model (b) we add control variables for differences in performance ("winning probability"), differences in self-assessment ("overconfidence") and risk-taking behavior ("risk").[12] With control variables, gender effects get smaller and are far from being significant in the word-order task (p = 0.491) but still significant in the summation task (p = 0.050). These results indicate that women do not generally shy away from competition. Instead, we find support for conjecture 1, i.e. that self-selection depends on the gender-connotation of the task.

4.3 Robustness Check: Stereotypes on Performance

In treatment IC-SUM and IC-WO, subjects guessed their relative rank given their stage 1 performance in groups of men only ("all-male") or women ("all-female") only. With this procedure we elicited perceived stereotypes about the performance of groups. Table 5 reports the average guessed rank by gender, task and reference group.

In the summation task both women (Wilcoxon signed rank test, p-value = 0.006) and men (Wilcoxon signed rank test, p-value = 0.002) belief on average that they are better in the all-female group than in the all-male group. This is an indicator for a strong stereotype that men are better than women in the mathematical task. In the word-order task, women think that they are better in the all male-group than in the all-female group while men think that they are slightly better in the all-female group than in the all-male group. However, these differences between group compositions are not significant

[12] For the winning probability, we bootstrap expected winning probability in the tournament given subjects' performance in the previous tournament stage. For the measure of overconfidence we used the difference between the optimal rank and the guessed rank in stage 1. For risk-taking behavior we insert a dummy variable for the choice of the tournament scheme in the lottery stage 5. Summary statistics of control variables are provided in appendix C.

Table 4: Linear probabilty model of tournament entry (IC/Stage 4)

Dependent variable	Tournament entry	
Column	(a) without controls	(b) with controls
SUM	0.265**	0.240*
	(0.117)	(0.141)
WO	0.234*	0.131
	(0.126)	(0.124)
SUM x female	-0.248**	-0.212**
	(0.107)	(0.107)
WO x female	-0.094	-0.073
	(0.103)	(0.102)
Winning probability		0.284*
		(0.170)
Overconfidence		0.049
		(0.043)
Risk		0.094
		(0.094)
Session dummies	Yes	Yes
Adj. R^2	0.22	0.23

Note: OLS regression with robust standard errors in parenthesis (n=128). Stars indicate levels of significance: * p<0.1, ** p<0.05, *** p<0.01

Table 5: Average guessed rank (IC/Stage 1)

Task	***Summation task***		***Word-order task***	
Given group	All-Male	All-Female	All-Male	All-Female
Male	2.6 (0.97)	2.0 (0.89)	2.7 (0.90)	2.6 (0.98)
Female	2.7 (0.93)	2.2 (0.95)	2.4 (1.05)	2.6 (0.91)
Total	2.7 (0.94)	2.1 (0.92)	2.6 (0.97)	2.6 (0.94)

Note: Standard errors in parentheses

(Wilcoxon signed rank test, p-value = 0.441 for male participants and 0.205 for female participants). These findings suggest, that task stereotypes reflected in the beliefs over relative performance are not that strong in the verbal task and differ by gender.

4.4 Robustness Check: Gender Stereotypes

Previously, subjects had to perform the task after their selection of incentive schemes. However, subjects' beliefs about expected performance in stage 4 may differ from previous performance in the tournament. We address this potential confound in stages 8 to 10 of the CO- and GC-TO-treatments where subjects submit their stage 1 performance to the different incentive schemes. Additionally, subjects select of incentive schemes given an all-male or all-female group.

Table 6 reports the choices of incent. In both tasks, less subjects chose the tournament when comparing a mixed (random) group to an all-male group. Likewise, more subjects tend to compete when comparing the all-female group with the mixed group. In the summation task, there is a large and significant gender gap in choices in the all-male group. This gap narrows in the mixed group and vanishes in the all-female group. As in the IC-SUM treatment, women tend to shy away from competition, however, this phenomenon is limited to competition against men or groups that potentially consist of men. In contrast to the choice in IC-WO, we also find a gender gap in the word-order task. This gap is largest in the all-female group while also significant in the all-male group and insignificant in the random group.

Table 7 reports the results of a linear probability model with the choice of the tournament scheme as the dependent variable. Columns (a), (c) and (e) replicate the descriptive results. In columns (b), (d) and (f), we add control variables for winning probability, overconfidence and risk as described in section 2.[13] With control variables, gender effects get smaller and are only significant in the all-male group (p = 0.018) for the summation task and in the all-female group (p = 0.022) and the all-male group (p = 0.044) for the word-order task.

Our results in the summation task support both, a stereotype that men perform better in the task (conjecture 2a) and that women tend to shy away from competing against men (conjecture 3). The results in the word-order task, however, are more puzzling. Here, women shy away from competing against both single-sex groups. One possible explanation might be that a stereotype about performance (conjecture 2b) and shying away from competition against men (conjecture 3) apply both but work in different directions.

[13] For the winning probability, we again bootstrap expected winning probability in the tournament given subjects' performance in stage 1 and the respective group composition. For the measure of overconfidence we used the difference between the optimal rank and the guessed rank in stage 2. For risk, we again used the choice of tournament in the lottery stage 5. Summary statistics of the control variables are reported in appendix C.

Table 6: Choice of payment regimes by group composition (stages 8-10)

Task	***Summation task***			***Word-order task***		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
All-male group (stage 8)						
Revenue sharing	25.0%	36.1%	30.6%	29.7%	34.2%	31.9%
Tournament	31.9%	11.1%	21.5%	27.0%	10.0%	18.8%
Fisher's exact test	0.009			0.030		
Mixed group (stage 9)						
Revenue sharing	25.0%	27.8%	26.4%	27.0%	31.4%	29.2%
Tournament	40.3%	22.2%	31.3%	31.1%	18.6%	25.0%
Fisher's exact test	0.049			0.239		
All-female group (stage 10)						
Revenue sharing	19.4%	18.0%	18.8%	31.1%	22.8%	27.1%
Tournament	44.4%	37.5%	41.0%	41.9%	21.4%	31.9%
Fisher's exact test	0.639			0.002		

Note: Residual category "piece rate" omitted

Table 7: Linear probability model of tournament entry (stages 8-10)

Dependent variable	Tournament entry					
Group	All-male		Mixed		All-female	
Column	(a)	(b)	(c)	(d)	(e)	(f)
SUM	0.354***	0.131	0.340***	0.099	0.285**	0.064
	(0.117)	(0.116)	(0.116)	(0.115)	(0.120)	(0.136)
WO	0.273**	0.045	0.253**	0.256**	0.417***	0.195
	(0.106)	(0.091)	(0.109)	(0.116)	(0.127)	(0.126)
SUM x female	-0.208***	-0.141**	-0.181**	-0.109	-0.069	-0.012
	(0.067)	(0.059)	(0.076)	(0.071)	(0.083)	(0.081)
WO x female	-0.171***	-0.126**	-0.131*	-0.083	-0.209***	-0.172**
	(0.063)	(0.062)	(0.072)	(0.072)	(0.076)	(0.074)
Winning probability		0.553***		0.514***		0.425***
		(0.109)		(0.096)		(0.097)
Overconfidence		0.075***		0.069***		0.054*
		(0.023)		(0.025)		(0.030)
Risk		0.076		0.146**		0.122*
		(0.058)		(0.067)		(0.070)
Treatment dummies	Yes	Yes	Yes	Yes	Yes	Yes
Session dummies	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.238	0.332	0.317	0.387	0.367	0.412

Note: OLS with robust standard errors in parentheses (n = 288). Treatment dummies for GC-TO-treatments. Stars indicate levels of significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$,

4.5 Robustness Check: Choice of Reference Groups

To further examine whether expectations of performance depend on gender and task, we examine the subject's choices of criteria in the GC-TO treatments. Table 8 reports the share of subjects choosing the respective attribute by task. Gender is the single most important attribute in both tasks for men and women (WO: 27.50%; SUM: 28.75% of all attribute choices). Additionally, there are significant differences in the choice of characteristics between tasks (Fisher's exact test = 0.049) that are driven by differences in the category gender and support gender stereotypes. In the summation task, only 3.75% chose male partners and 25% chose female partners, in the word order task both sexes were chosen with equal probability of 13.75%. This result suggests that a strong stereotype in the summation task does also affect the choice of groups as described in conjecture 4a and that the word-order task does not evoke specific gender associations which supports our conjecture 4b.

Table 9 reports the choices of group compositions given stage 1 performance and the tournament or weakest-link in the IC-treatments. In the summation task male competitors are chosen in 7.81% of all cases while this fraction increased to 18.75% in the word-order task. The proportion of women competitors decreased from 48.44% in the summation task to 29.69% in the word-order task. A Fisher's exact test confirms that these differences are significant between the two tasks (p-value: 0.045). These choices confirm the findings reported in table 8, giving further support that choices are driven by gender-task stereotypes.

Choices in the weakest-link game in the summation task are not as pronounced but confirm this picture: over 42% of the subjects chose male partners while only 23% chose female partners. Similar, but even less pronounced patterns can be found in the word-order task where male partners are slightly preferred to female partners.

5 A Field Experiment

5.1 Experimental Design and Procedure

In order to test the validity of our results in a broader population, we conducted a framed field experiment (See Harrison and List, 2004, for a definition) with subjects in an age range from 6 to 67 years. We applied two different treatments that varied the task and whithin each treatments subjects could either select into a competitive environment or chose to be paid by peice. In the first treatment we applied the word-order task described previously and a ball throwing task that was applied by Gneezy et al. (2009) to detect differences in competitiveness between men and women in matrilineal and patrilineal societies. The experiment was conducted during the "Long Night of Science" at the University of Jena, an event where all faculties display their current research to a

Table 8: Chosen criteria by gender (GC-TO/Stage 4)

Treatment	**GC-TO-SUM**			**GC-TO-WO**		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
	%	%	%	%	%	%
No choice	*25.00*	*20.00*	*22.50*	*7.50*	*22.50*	*15.00*
Gender	*22.50*	*35.00*	*28.75*	*27.50*	*27.50*	*27.50*
Gender: male	5.00	2.50	3.75	17.50	10.00	13.75
Gender: female	17.50	32.50	25.00	10.00	17.50	13.75
Age	*7.50*	*7.50*	*7.50*	*20.00*	*22.50*	*21.25*
Age: below	5.00	5.00	5.00	7.50	7.50	7.50
Age: above	2.50	2.50	2.50	12.50	15.00	13.75
Siblings	*5.00*	*10.00*	*7.50*	*7.50*	*7.50*	*7.50*
Siblings: below	0.00	2.50	1.25	2.50	0.00	1.25
Siblings: above	5.00	7.50	6.25	5.00	7.50	6.25
Distance	*2.50*	*7.50*	*5.00*	*20.00*	*2.50*	*11.25*
Distance: below	0.00	2.50	1.25	10.00	0.00	5.00
Distance: above	2.50	5.00	3.75	10.00	2.50	6.25
Sport	*27.50*	*17.50*	*22.50*	*10.00*	*15.00*	*12.50*
Sport: below	10.00	10.00	10.00	5.00	5.00	5.00
Sport: above	17.50	7.50	12.50	5.00	10.00	7.50
Societies	*10.00*	*2.50*	*6.25*	*7.50*	*2.50*	*5.00*
Societies: below	5.00	2.50	3.75	2.50	2.50	2.50
Societies: above	5.00	0.00	2.50	5.00	0.00	2.50
Differences by gender		Pearson χ^2	p-value		Pearson χ^2	p-value
Attributes (6)		5.665	0.462		9.903	0.129
Characteristics (12)		8.822	0.718		14.394	0.276
Differences by treatment						
Attributes (6)		10.803	0.095			
Characteristics (12)		19.848	0.070			

broader audience. The experiment was presented as an attraction for the visitors where they could win prizes in the form of sweet or sour candies. Subjects were welcomed at the reception desk and rolled a six-sided dice to determine their task (1-3 for word order and 4-6 for ball throwing). They then obtained a game card, where their decisions and results were recorded and the tasks were explained.[14] Before subjects chose the payment regime and performed the task, they performed in a 90-second practice round.

The tasks

In the word-order task subjects were asked to order five words to build a grammatically correct sentence. The task was presented as described in section 2, but performed with paper and pencil. Subjects had to complete as many sentences as possible until 90 seconds had passed.

In the other task, subjects were asked to throw soft tennis balls in a bucket which was 3 meters away. In the practice phase, subjects had 10 tennis balls to throw into the bucket. In the performance phase, they then had 90 seconds to throw as many tennis balls as possible into the bucket.

Choice of payment regime

After the practice round, subjects could choose how they would like to be payed in the performance phase. Subjects had the choice between a piece rate or a competitive scheme. In the piece rate, subjects earned 1 point per completed task. In the competitive regime the subject's performance is compared with a randomly drawn result performed five days earlier in a pilot study by other participants. If the subject at least matched the result of her opponent, she got 2 points per task. If she had less, she got 0.25 points per solved task. Subjects chose the payment regime in private by ticking the respective box on their game card. Earned points were exchanged for candies at an exchange rate of 1:2 after the experiment.

Controls

As in the laboratory experiment, we used the question for the willingness to take risk from the GSOEP (Wagner et al., 2007). Before the practice round, subjects were asked to guess their performance in this round and received an additional gift for guessing correctly. This resulted in a measure of task-specific overconfidence.[15] As we do not know whether competitiveness is age-dependent, we asked for the age of the participant

[14] Translated copies of the game card can be found in the appendix.

[15] We are aware that this compensation theoretically may not elicit the mean of the distribution of a subject's beliefs, but will favor modal predictions. We chose this mechanism because it is easier to explain

in a short post-experimental questionnaire. In the analysis, we exclude all subjects aged below 14, because we wanted to make sure that the risk question is understood properly.[16] We report summary statistics of the controls in appendix D.2.

5.2 Results

Table 13 shows summary statistics of selected payment regimes. Men chose to compete more often in the ball task and women chose to compete more often in the verbal task. A Fisher's exact test shows that choices are significantly dependent on gender in the ball task (Fisher's exact test, p-value: 0.050) but not in the word-order task (Fisher's exact test, p-value: 0.574). However, as we have a lack of balance in age and a higher performance of women in the verbal task, we introduce controls to check the robustness of these results.

Table 11 reports the results of linear probability models using task dummies, interactions of task and gender dummies and additional control variables. Column (1) reproduces the results of the nonparametric tests. Women chose competition significantly less in the ball task and chose to compete more often than men in the word-order task. Controlling for age in column (2), we find that older subjects compete more often. The point estimate of the coefficient on $ball \times female$ gets smaller but remains positive.

6 Conclusions

We offer an explanation for differences in sorting into competitive environment based on the perception of the underlying task. We find little impact of gender on performance in competitive environments or that women enter the tournament systematically less often. Given the evidence of the experiments, we conclude that selection into competitive environments interacts non-trivially with the underlying task and gender. Gender stereotypes in the mathematical task indicate that men are expected to perform much better than women and drive gender differences in performance and the choice of competitive environments. This alters the interpretation of the results by previous studies that claimed that women are less competitive *per se*. Consistent with the previous literature (e.g., Niederle and Vesterlund, 2007), we find gender differences in self-selection of women into competition in the mathematical task, even after controlling for winning probability, self-assessment, and risk attitudes. As we do not find these differences in the verbal task, we interpret our results differently. We reject that women's choices can only

than a more complex rule such as quadratic scoring.Offerman et al. (2001) offer evidence suggesting that even a flat fee leads to good judgments; hence we do not expect the choice of scoring rule to matter much.

[16] This measure was chosen, because in Germany individuals can be held legally responsible for their actions from the age of 14 (see §19 StGB [German Criminal Code]).

Table 9: Choice of gender-groups in IC-treatment (stages 8-10)

Incentive Scheme	*Tournament*					
Treatment	**IC-SUM**			**IC-WO**		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
All-male group	13.3%	2.9%	7.8%	15.6%	21.9%	18.8%
All-female group	46.7%	50.0%	48.4%	34.4%	25.0%	29.7%
Differences by task	Fisher's exact test					
Male:	0.653					
Female:	0.020					
Total:	0.045					
Incentive Scheme	*Weakest-link*					
Treatment	**IC-SUM**			**IC-WO**		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
All-male group	43.3%	41.2%	42.2%	34.4%	25.0%	29.7%
All-female group	30.0%	17.7%	23.4%	21.9%	18.8%	20.3%
Differences by task	Fisher's exact test					
Male:	0.370					
Female:	0.355					
Total:	0.204					

Note: The residual category "random group" is omitted.

Table 10: Choice of payment regime

Task	***Ball task***			***Word-order task***		
Group	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
Tournament	61.54%	30.77%	46.15%	45.00%	55.88%	51.85%
Num. obs:	26	26	52	20	34	54
Fisher's exact test	0.050			0.574		

Note: The residual category "piece rate" is omitted.

Table 11: Choice of competition

Dependent variable	Choice of tournament			
Column	(1)	(2)	(3)	(4)
	Treat.	Age	Practice	Risk
Word	0.45***	-0.056	-0.42	-0.79
	(0.11)	(0.43)	(0.47)	(0.51)
Ball	0.62***	0.090	-0.17	-0.57
	(0.097)	(0.44)	(0.46)	(0.50)
Word × female	0.11	0.11	0.029	0.050
	(0.14)	(0.14)	(0.14)	(0.14)
Ball × female	-0.31**	-0.31**	-0.28**	-0.25*
	(0.13)	(0.14)	(0.14)	(0.13)
Age		0.035	0.035	0.038
		(0.024)	(0.024)	(0.025)
Age, squared		-0.00050*	-0.00047	-0.00052*
		(0.00029)	(0.00029)	(0.00029)
Result practice			0.10**	0.10**
			(0.041)	(0.039)
Overconfidence			0.00023	-0.0024
			(0.031)	(0.032)
Risk				0.064**
				(0.025)
Adj. R^2	0.500	0.506	0.541	0.568

Note: OLS with robust standard errors in parentheses (n = 106). Stars indicate levels of significance: * p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01,

be explained by attitudes toward competition (against men) *per se*. Instead, we find that gender-task stereotypes are an important confounding factor which additionally explains remaining gender differences in self-selection.

However, these different interpretations of behavioral differences in experimental settings are important as they affect how firms and policy makers can provide better opportunities for women. Many aspects of stereotypes certainly affect behavior in very subtle, often unrecognized ways. A better understanding for the emergence of stereotypes is therefore necessary. Nevertheless, gender-job associations based on false or exaggerated stereotypes allow for different policy recommendations that alter stereotypes: for example, to avoid teaching material that transports gender-specific stereotypes as reported for educational software for pre-schoolers by Sheldon (2004) where technical professions were mainly illustrated by male characters.

Our findings also yield interesting insights regarding the use of real effort tasks in economic experiments. The results show that real effort tasks are frames that are not neutral with respect to stereotypes. These stereotypes may or may not be justified by actual differences in performance. Expectations about the performance of subgroups in the experimental subjectpool are an important confounding factor that needs to be controlled for or minimized by more neutral tasks.

Fruitful further research will have to determine sources of stereotypes which is the basis for alleviating the negative economic consequences resulting of those misperceptions. Furtherermore, one has to assess the role of information has on subjects, as there is an ongoing debate, as the effect could either increase the effect of an already existing stereotype (as has been argued by Wheeler and Petty, 2001 and others cause by what the literature called *stereotype threat*) or – by making it salient – help to alleviate them (for experimental evidence see Johns et al., 2005).

A Instructions

Printed intructions (English translation)

Welcome to this experiment and thank you for your participation!

In this experiment - financed by the Deutschen Forschungsgemeinschaft - you can earn money, depending on your own performance and decisions. Therefore, it is important that you read the instructions carefully. If you have any questions during the experiment, please raise your hand. We will then come to you and answer your question. Please pose your question quietly so that others cannot hear. All participants of this experiment receive the same instructions. The information on the screen, however, is private, so please do not look at the screens of other participants and do not talk to each other. If you do not stick to these rules, we unfortunately have to exclude you from the experiment. Please switch off your mobile phones now.

General schedule

This experiment lasts about 90 minutes. This experiment consists of one practice task, 10 tasks, two bonus questions and a questionnaire.

Before each task, you receive detailed instructions on the task and the payment mechanisms according to which the task is paid. Please read the instructions carefully. In the practice task and in six out of the 10 other tasks, you are asked to perform. After completing the task, you are informed how many correct solutions you have come up with. Then, please enter then the name of the task and the number of correct solutions in the table on the back of these instructions. You are reminded that no technical aids (mobile phones, calculators, computers, etc.) are allowed. If this rule is not followed, you will be expelled from the experiment! After the last task, you are asked to answer five [two] additional bonus questions. In these questions, you can earn €0.5 [€1] for each correct answer independent of your performance in the tasks.

Payments

At the end of the task part, one task will be randomly chosen for payment. For this purpose, a volunteer will draw a tennis ball with a number on it from an urn. This number determines the task for payment. Independent of the chosen task, you will receive the payment for the bonus questions. You will be paid after the questionnaire so that no other participant will learn how much you earned.

Further schedule

After you have read the instructions carefully, please wait for the other participants and then start with the computer program on your screen. Please do not forget to enter the data in the table on the back.

Good luck!

B Questionnaire

Do you have siblings?
If yes: how many brothers? And how many sisters?

Schooling

When you think back: how strongly were your parents were interested in your school performance?

Which grades did you obtain at school in your final examinations in the following three subjects?
German, Mathematics, First foreign language:

Was your school: single-sex/mixed-sex?

Attitudes

Are you in general a person who is impatient or very patient?

Are you in general a person who likes to take risks or tries to avoid risks?

Are you in general a person who takes time and thinks before acting or are you an impulsive person?

Stereotypes

See Osgood et al. (1957) for the original questions and Stier (1999) for a German translation.

Competitive behavior

Have you practiced music in your youth like singing or playing an instrument?

Do you actively participate in sports?

Which is the most important sport you practice?

How old were you when you started this sport?

Where and with whom do you practice this sport?

Do you compete in this sport?

Work attitudes

For your choice of work, different things may be important to you; please list in the following points which are important to you:
Secure job, High salary, Career prospects, A job that is respected by others, A job that leaves you free time, An interesting job, A job with self-responsibility, Contact to other people, A job that is important for society, Secure and healthy environment, Enough time for family, A job in which one can help others

C Controls

Table 12: Statistics of control variables

Task	*Summation*			*Word-order*		
Group	**Male**	**Female**	**T-test**	**Male**	**Female**	**T-test**
Only IC/Stage 4						
Winning prob	0.379	0.222	**	0.350	0.336	
	(0.336)	(0.245)		(0.238)	(0.262)	
Overconfidence	-0.033	0.265		-0.406	-0.219	
	(0.964)	(0.963)		(0.875)	(1.211)	
Risk	0.300	0.206		0.375	0.063	***
	(0.466)	(0.410)		(0.492)	(0.043)	
Stages 8-10						
Win prob (all-male)	0.275	0.210		0.282	0.231	
	(0.293)	(0.231)		(0.299)	(0.269)	
Win prob (random)	0.315	0.249		0.302	0.255	
	(0.322)	(0.268)		(0.315)	(0.287)	
Win prob (all-female)	0.361	0.296		0.316	0.273	
	(0.353)	(0.308)		(0.327)	(0.302)	
Overconfidence	0.403	0.097		-0.054	-0.157	
	(1.109)	(1.189)		(0.935)	(0.958)	
Risk	0.319	0.208		0.270	0.174	
	(0.470)	(0.409)		(0.447)	(0.380)	

Notes: Mean reported. Standard deviation in parenthesis. Stars indicate significant differences between male and female using a two-sided t-test with unequal variances: * $p<0.1$, ** $p<0.05$, *** $p<0.01$

D Field Experiment

D.1 Game cards

Game Card Throwing Ball

Description of the game

In this game you have to throw balls into a bucket. You start with a trial round, where you have 10 balls of which you are supposed to throw as many as possible into the bucket.

Before you play the same game in the main round, you have to decide for a payment mode. This decision you will take alone and privately. The assistant at the game will help you if you have further questions. The game in the main round starts with "START" and ends with "STOP".

First you have to guess how many balls you will manage to throw into the bucket in the trial round.

Estimate trial round: _____

Balls in bucket in trial round _____

Please select a payment mode:

Modus A: You will get per ball in the bucket one point

Modus B: You will play against a randomly chosen person who has done this task before. The number of balls in the bucket the other person had will be drawn from an urn after you have completed the task. If you have more or the same amount of balls in the bucket than your opponent, you will get 2 points per ball in the bucket. Otherwise you get 0.25 points per ball in the bucket.

Selected payment mode: _____

Balls in bucket in main round: _____

Game Card Sentences

Description of the game

In this game you have to unscramble 5 words so that they make up one sentence. You begin with a trial round in which you have 45 seconds to build as many correct sentences as possible. The assistant takes the time. At the word "START" the words will be uncovered and you can work at "STOP" the words will be covered again.

Before you play the same game in the main round, you have to decide for a payment mode. This decision you will take alone and privately. The assistant at the game will help you if you have further questions. The game in the main round starts with "START" and ends with "STOP".

First you have to guess how many balls you will manage to throw into the bucket in the trial round.

Estimate trial round: _____

Solutions in trial round _____

Please select a payment mode:

Modus A: You get 1 point per correct sentence.

Modus B: You will play against a randomly chosen person who has done this task before. The number of correct sentences the other person had will be drawn from an urn after you have completed the task. If you have more or the same amount of correct sentences than your opponent, you will get 2 points per correct sentence. Otherwise you get 0.25 points per correct sentence.

Selected payment mode: _____

Solutions in main round: _____

D.2 Controls

Table 13: Summary statistics of characteristics of subjects aged above 14

	Age	*Female*	*Like doing task*	*Risk*	*Overconfidence*
Word order					
Mean	29.6	63%	6.5	5.6	0.9
Std.dev.	12.6	0.49	2.5	2.1	1.9
Median	25	1	7	5	1
Maximum	50	1	10	8	3
Minimum	19	0	2	3	-1
Ball					
Mean	30.1	50%	7	6.0	0.9
Std.dev.	11.1	0.5	2.6	1.8	2.2
Median	26	0.5	8	6	1
Maximum	47	1	10	8	3
Minimum	20	0	3	4	-1
Total					
Mean	29.8	57%	6.8	5.8	0.9
Std.dev.	11.8	.50	2.5	1.9	2.0
Median	26	1	7	6	1
Maximum	50	1	10	8	3
Minimum	20	0	3	3	-1

References

Akerlof, George A. and Kranton, Rachel E. (2000): "Economics and Identity", *Quarterly Journal of Economics*, 115(3), 715–753.

Barber, B. M. and Odean, T. (2001): "Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment", *Quarterly Journal of Economics*, 116(1), 261–292.

Benjamin, D. J., Choi, J. J., and Strickland, J. (2007): "Social Identity and Preferences", NBER Working Paper No. 13309.

Booth, Alison L. (2009): "Gender and competition", *Labour Economics*, 16(6), 599–606.

Camerer, C. and Lovallo, D. (1999): "Overconfidence and Excess Entry: An Experimental Approach", *American Economic Review*, 89(1), 306–318.

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

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G.G. (2005): "Individual Risk Attitudes: New Evidence from a Large Representative, Experimentally-Validated Survey", IZA Discussion Paper No. 1730.

Dohmen, Thomas and Falk, Armin (2010): "Performance Pay and Multidimensional Sorting - Productivity, Preferences and Gender", *American Economic Review*, Forthcoming.

Eagly, Alice H. and Karau, Steven J. (1991): "Gender and the emergence of leaders: A meta-analysis", *Journal of Personality and Social Psychology*, 60(5), 685–710.

Eagly, Alice H. and Karau, Steven J (2002): "Role congruity theory of prejudice toward female leaders", *Psychological Review*, 109(3), 573–598.

Eckel, C. C. and Grossman, P. J. (2008): "Differences in Economic Decisions of Men and Women: Experimental Evidence", in *Handbook of Experimental Results*, edited by Plott, C. R. and Smith, V. L., Amsterdam: North-Holland, vol. 1, pp. 509–519.

Fischbacher, U. (2007): "Zurich toolbox for readymade economic experiments", *Experimental Economics*, 10, 171–178.

Gneezy, U., Niederle, M., and Rustichini, A. (2003): "Performance in Competitive Environments: Gender Differences", *Quarterly Journal of Economics*, 118(3), 1049–1074.

Gneezy, Uri and Rustichini, A. (2004): "Gender and Competition at a Young Age", *American Economic Review, Papers and Proceedings*, 94(2), 377–381, ISSN 00028282.

Gneezy, Uri, Leonard, Kenneth L., and List, John A. (2009): "Gender Differences in Competition: Evidence from a matrilineal and a patriarchal society", *Econometrica*, 77(5), 1637–1664.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments", *GWDG Bericht*, 63, 79–93.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008): "Culture, gender, and math", *Science*, 320(5880), 1164–1165.

Harrison, Glenn W. and List, John A. (2004): "Field Experiments", *Journal of Economic Literature*, 42(4), 1009–1055, ISSN 00220515.

Johns, M., Schmader, T., and Martens, A. (2005): "Knowing Is Half the Battle", *Psychological Science*, 16(3), 175–179.

Lundeberg, M. A., Fox, P. W., and Puncochar, J. (1994): "Highly confident but wrong: gender differences and similarities in confidence judgments", *Journal of Educational Psychology*, 86(1), 114–121.

Marmaros, D. and Sacerdote, B. (2006): "How do Friendships Form?", *Quarterly Journal of Economics*, 121(1), 79–119.

Mayer, A. and Puller, S. (2008): "The old boy (and girl) network: Social network formation on university campuses", *Journal of Public Economics*, 92, 329–347.

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

Niederle, M., Segal, C., and Vesterlund, L. (2008): "How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness", NBER Working Paper No.13923.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002): "Math = male, me = female, therefore math $\neq$ me", *Journal of Personality and Social Psychology*, 83(1), 44–59.

OECD (2006): "Education policy analysis: focus on higher education 2005-2006", Paris: Organisation for Economic Co-operation and Development.

Offerman, Theo, Sonnemans, Joep, and Schram, Arthur (2001): "Expectation Formation in Step-Level Public Good Games", *Economic Inquiry*, 39(2), 250–69.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957): *The Measurement of Meaning*, Urbana: University of Illinois Press.

Phelps, Edmund S. (1972): "The Statistical Theory of Racism and Sexism", *American Economic Review*, 62(4), 659–661.

Schwieren, C. and Weichselbaumer, D. (2009): "Does competition enhance performance or cheating? A laboratory experiment", *Journal of Economic Psychology*, ISSN 0167-4870, forthcoming.

Sheldon, J. P. (2004): "Gender Stereotypes in Educational Software for Young Children", *Sex Roles*, 51(7-8), 433–444.

Steele, C. (1998): "Stereotyping and its threat are real", *American Psychologist*, 53(6), 680–681.

Steele, Jennifer R., Reisz, Leah, Williams, Amanda, and Kawakami, Kerry (2007): *Women and minorities in science, technology, engineering and mathematics*, Edward Elgar Publishing, chap. Women in Mathematics: examining the hidden barriers that gender stereotypes can impose, ISBN 1845428889, 9781845428884, pp. 159–183.

Stier, W. (1999): *Empirische Forschungsmethoden*, Berlin: Springer.

Van Lange, P., Otten, W., De Bruin, E., and Joireman, J. (1997): "Development of Prosocial, Individualistic, and Competitive Orientations: Theory and Preliminary Evidence", *Journal of Personality and Social Psychology*, 73(4), 733–746.

Wagner, G. G., Frick, J. R., and Schupp, J. (2007): "The German socio-economic panel study (SOEP)–scope, evolution and enhancements", *Schmollers Jahrbuch*, 127(1), 139–169.

Wheeler, S.C. and Petty, R.E. (2001): "The effects of stereotype activation on behavior: A review of possible mechanisms", *Psychological Bulletin*, 127(6), 797–826.