The Korean Economic Review
Volume 33, Number 1, Summer 2017, 95-126.

Gender Interaction in Teams: Experimental Evidence on Performance and Punishment Behavior

SeEun Jung* · Radu Vranceanu**

This paper reports the results from an experiment where men and women are paired to form a two-member team and asked to execute a counting task. An individual's payoff is proportional to the joint production of right answers. Participants who perform better than their partner in the task can punish him or her by imposing a fine. We manipulate the pairs' gender compositions to analyze whether an individual's performance and sanctioning behavior depend on his or her gender and the gender of his or her partner, which is revealed to the subjects at the beginning of the experiment. The data show that, conditional on underperformance, women are sanctioned more often and more heavily than men; however, if they are sanctioned, men tend to improve their performances, while women's performances do not change.

JEL Classification: M52, J16, C91
Keywords: Gender, Real-effort task, Team production, Performance, Punishment, Discrimination

I. Introduction

Because in many circumstances individual contributions to the production process can be difficult to measure, companies often choose to pay team members proportional to the teams' output (Lawler and Mohrman, 2003; Boning et al., 2007). This compensation scheme creates an environment that is favorable to free riding. Indeed, if every team member rests upon the efforts of his or her partners, the game presents an inefficient Nash equilibrium where all of the team members shirk. Several scholars have argued that peer monitoring and peer pressure, which are

Received: Sept. 11, 2016. Revised: Dec. 2, 2016. Accepted: Jan. 23, 2017.

* First Author, Assistant Professor, Department of Economics, Inha University, Incheon, South Korea. E-mail: jse@inha.ac.kr

** Corresponding Author, Professor, Economics Department, ESSEC Business School and THEMA. PB 50105, 95021 Cergy, France. E-mail: vranceanu@essec.fr

understood as sanctions that are taken by the group to punish individuals who deviate from the cooperative strategy, can help to alleviate the free-riding problem.[1] Such sanctions need not be monetary; mockery, intimidation and the social exclusion of those who deviate from the group's norm are ubiquitous phenomena in modern organizations (Fehr and Gächter, 2000). However, beyond the former "discipline" effect, sanctions can entail undesired outcomes because they generate negative emotions that may harm performance. Psychologists refer to this negative outcome as the "motivation crowding-out" effect (for a survey, see Frey and Jegen, 2001 and Festré and Garrouste, 2014).[2]

One important characteristic of production in teams is gender composition. Depending on the task and the incentive scheme, men or women can perform better, at least as far as we compare their average performance on the task. Nevertheless, for a given task and incentive scheme, the relative performance of men and women can vary depending on the gender with which they are paired. An important research question in the contemporary context where women's participation in the labor force is continuously rising is whether gender interactions can explain variations in individual performance.

To analyze how the gender composition of a team affects the performance and sanctioning behavior of individuals, we use the real-effort experiment that was developed by Mohnen et al. (2008) and extended by Vranceanu et al. (2015) to allow for explicit peer punishment. At the outset of the experiment, the subjects are allocated to production pairs. The real task participants must execute tasks that consist of counting the 7s in blocks of figures, which are successively displayed on a computer screen during six successive rounds of four minutes each. In each round, individuals receive an equal share of the team's output, which is proxied by the total number of correct answers.[3] Hence the benefit of each player increases with the effort of the other player. We use the same payoff calibration as Vranceanu et al. (2015) such that without punishment individuals have an incentive to free ride (they are paid to rest). The best performer on a team has the option to punish the less productive partner by applying a monetary sanction, which entails a cost to the punisher. Because we want to study the response to sanctions in a longer-term relationship, which is closer to the team production in firms, the experiment uses a partner design whereby the teams are kept invariant for the six rounds of a given

[1] See for instance: Alchian and Demsetz (1972), Holmstrom (1982), McAfee and McMillan (1991), Itoh (1991), Kandel and Lazear (1992), Legros and Matthews (1993), Barron et al. (1997).

[2] This motivation crowding-out effect has already been observed in experiments that test the principal/agent game (Dickinson and Villeval, 2008; Kirstein, 2008), the trust game (Fehr and List, 2004; Fehr and Rockenbach, 2003; Houser et al., 2008) or the ultimatum game (Gneezy et al., 2003).

[3] Note that this compensation scheme should encourage cooperation, as in the experiment by Kuhn and Villeval (2015), and it differs from the competitive environment that is specific to the experiments of Gneezy et al. (2003) or Niederle and Vesterlund (2007, 2011), in which individual payment is based on a tournament.

session. As an original development of this paper, we create mixed gender and homogenous gender pairs. The players are informed of the gender of the partner, while the anonymity of the participants is preserved. Thus, this experiment seeks to detect the "genuine gender effect", independent of any emotions or social stances that would develop if the partners could establish eye contact or communicate. A subject cannot participate in more than one experimental session in a typical between-subjects design.

This experiment will allow us to address three key research questions:

(1) Does individual productivity change depending on the gender of the teammate? Further, do observed differences depend on differences in ability or in work organization, sanctioning patterns and different responses to sanctions?

(2) Do subjects adapt their sanctioning behavior to the gender of their opponent? (If so, does this behavior vary across male and female subjects?)

(3) Is there a gender difference in the response to sanctions?

In essence, our results indicate that, in this task, individuals in homogenous gender teams tend to perform slightly better than individuals who are placed in mixed gender teams. These differences in individual productivity cannot be explained by direct gender interactions; however, they are related to the different responses of males and females to sanctions. If men and women have similar sanctioning behaviors, when they under-perform, women are sanctioned more frequently than men. In turn, when men are sanctioned, men's performance improves, whereas women's performance does not, as if, in the case of women, the motivation crowding out effect will offset the "discipline" effect.[4]

An early attempt to detect the gender effect in a real-effort laboratory experiment (solving mazes) with revenue sharing was provided by Ivanova-Stenzel and Kübler (2011). They analyzed individual behavior in a no-sanction, team production game with three types of teams - all men, all women and mixed. They found no significant difference between the individual productivity of men and women in single-sex teams; however, in mixed teams, men solve significantly more mazes than their female partners, which is a result that they explain by gender stereotypes according to which men should support women.

Some recent empirical studies have shown that the gender composition of a team may have an impact on the team's performance. For instance, Apesteguia et al. (2012) analyze information from a large database on a well-known business game that is played by self-selected teams of three students (*StratX-l'Oréal*). They show that, even when controlling for personal characteristics, all-women teams perform worse than all-men and mixed teams. A similar result is obtained by Lamiraud and Vranceanu (2015) using panel data from a different business game (*Kalystée-l'Oréal*)

[4] We also used the dataset of Vranceanu et al. (2015) to confirm that there is no gender effect in groups that perform a similar real-effort task without knowing the gender of the partner.

with a random allocation of students to teams of five. In their study, the all-men and mixed teams with a majority of women perform the best. Hoogendoorn et al. (2013) collect data on small businesses that are created and operated for one academic year by teams of students who are enrolled in the entrepreneurship program of the Amsterdam College of Applied Sciences, and they find that teams with an equal gender mix perform better than male-dominated teams in terms of profits and sales.

Other experimental papers analyze how gender interactions affect individual behavior in two-person games. In the Dictator game, one participant receives an endowment and can share it with an anonymous partner, knowing that the latter can take no subsequent action. Thus far, there is no clear evidence on gender effects in such games. Eckel and Grossman (1998) and Fehr et al. (2006) report that women tend to be more generous than men, while Frey and Bohnet (1995) or Carpenter (2007) find no gender differences in giving. With respect to gender interactions, Ben-Ner et al. (2004) found that women give systematically less to women than to men and persons of an unknown gender. Finally, Dufwenberg and Muren (2006) show that women receive more favorable treatment than men in the Dictator game.

Differences in how men (women) react when they are paired with the same (opposite) gender partners have been observed in standard Ultimatum games. Eckel and Grossman (2001) find that women make more generous offers than men, that offers that are made by women do not depend on the partner's gender, and that they are more likely to be accepted. Solnick (2001) finds that women and men make similar offers; however, women are more demanding than men when the opposer is a woman, and they are less demanding when the opposer is male. In a 2x2 design male/female and origin of Jewish immigrants to Israel (Ashkenazic / Sephardic), Fershtman and Gneezy (2001) found that men (but not women) discriminated against Ashkenazic men (but not against women), offering them less than they did to Sephardic men. Sutter et al. (2009) study a Power-to-take game (which can be seen as a "reverse Ultimatum game"), and they show that "take authorities" demand more from responders of the same gender, and responders' destruction rates are higher when the take authority has the same gender.

Several scholars have argued that observed differences in wages and promotions can be related to differences in bargaining behavior. Some contextual experiments tend to back this assumption. For instance, Bowles et al. (2007) provide lab-based evidence that shows that women are less inclined than men to negotiate a pay raise when the decision-maker is a man, and they explain this outcome by increased nervousness. Hederos Eriksson and Sandberg (2012) find that men are more prone than women to initiate a pay rise negotiation if the counterpart is a woman but not when the counterpart is a man. Dittrich et al. (2014) develop a face-to-face wage negotiation experiment and show that males who play the role of employers pay lower wages to female employees than female employers pay to male employees.

To our knowledge, there are few experimental studies that focus on gender interactions in sanctioning behavior. Using experimental evidence from a public good game that was played with Ghanaian subjects, Asiedu and Ibanez (2014) found no significant gender differences in punishment by "monitors" of players who deviate from cooperation in the patriarchal society. However, male monitors tend to sanction more often than female monitors in the matrilineal society, which is explained by the authors in terms of the use power to counterbalance the higher status of women. Finally, there can be a gender effect in the response to a sanction, and in particular the extent of the motivation crowding-out mechanism. In an interesting field study, Rask and Tiefenthaler (2008) analyzed why relatively fewer women decide to major in economics, and they found that their decision to drop such classes is motivated by the poor grades that they received in introductory courses in economics, whereas men are less discouraged by the negative signal that was provided by the same poor grades.

The paper is organized as follows: The next section introduces the experimental design. Section 3 presents the main results. The final section presents our conclusions and provides some managerial implications, with all of the caveats that are related to the challenge of extrapolating from such simple experiments. The conclusions are deemed to be relevant for managers in "nuts and bolts" firms.

II. Experimental Design

All of the subjects were recruited from the student population of the ESSEC Business School (France), and they had answered an advertisement for paid decision experiments.[5] Six sessions were organized at the ESSEC Experimental Lab with a total of 132 subjects, as is summarized in Table 2. At the beginning of the experiment, the subjects are matched in pairs at random. Each session comprises six identical four-minute rounds. A team composition is not changed across rounds. Interaction is anonymous, and hence the subjects do not know who their partners are. They play the game on a computer screen, and they cannot establish eye contact with one another. The instructions (provided in the Appendix) and data collection are computerized. The program was developed using z-Tree (Fischbacher, 2007).[6] The payoffs are denominated in Experimental Currency Units (ECU), with an exchange rate of 100 ECU = € 2.5.

In a typical round, the effort task as was used in Mohnen et al. (2008); Pokorny

[5] As "Grande Ecole" students, this group is relatively homogenous in terms of computing and intellectual abilities, age and educational background. It should be acknowledged that students are admitted to ESSEC after succeeding in a competitive national exam, with a demanding test in mathematics.

[6] The computer program was developed by Delphine Dubart at the ESSEC Experimental Lab.

(2008); Vranceanu et al. (2015), is as follows. The subjects are asked to count the number of 7s in blocks of random numbers that are successively displayed on the computer screen during a period of four minutes. The typical block has 30 columns and 6 rows (see Appendix). In each block, the number of 7s varies at random between 11 and 24, with an average of 18. This task is of interest for experimental research because it does not require any particular skill or computing ability. Note that the difficulty of the task depends on the total number of figures in a block (180), and it does not depend on the number of 7s; thus it should not vary from one block to another. At the beginning of the round, the computer displays the first block. When the participant finishes counting, he or she indicates the number of 7s in a box and then clicks "validate"; the computer records the answer (and checks it), and then it displays another block of figures. Thus, the number of blocks that are displayed during the four minutes of a round depends on the speed of the participants counting. For each round, a player's reward for work is calculated according to a simple rule (1). Let us consider the pair that is composed of players i and j. Let N_{it} and N_{jt} be the numbers of correct answers that are provided by the individuals $(i\ j)$, respectively, in round t. The reward "from work" is a linear function in the sum of correct answers that are provided by the two players $(N_{it}+N_{jt})$.

Instead of counting, the participant has the alternative to rest; if he presses the button "time-out", the computer cancels the count task by displaying a neutral screen (ESSEC logo) for 20 seconds.[7] If during the round t the participant i presses the time-out button k_{it} times, he or she also obtains an extra $6k_{it}$ ECU. This option can be viewed as an opportunity cost of working.

Finally, the player who out-performed (i.e., provided more correct answers than the other) is asked whether he or she wishes to impose a penalty on his or her partner. We decided to forbid punishment by persons who underperformed to rule out a retaliation motive or spiteful behavior. If he or she answers "yes", he or she can penalize the other player with an amount p, with $p \in [1;30]$ ECUs. Punishment is costly: each unit of sanction entails a cost of 0.30 ECUs for the punisher. Such a linear punishment technology, which involves a constant cost per unit of punishment, has been used in many other studies (Fehr and Gächter, 2000; Falk and Fischbacher, 2005; Nikiforakis and Normann, 2008). In the event of equal performance, players are not given the penalty option.

With these rules, for each round t, compensation functions for each player (i,j), can be written in the case when $N_{jt} > N_{it}$ as:

[7] The time-out button is deactivated 20 seconds before the end of the round (this is the average time that is needed to count the 7s in the last block of numbers).

$$\begin{cases} Z_{it} = 10\dfrac{N_{it}+N_{jt}}{2} + 6k_{it} - f_{jt}p_{jit} \\ Z_{jt} = 10\dfrac{N_{it}+N_{jt}}{2} + 6k_{jt} - 0.3f_{jt}p_{jit} \end{cases} . \quad (1)$$

where f_{jt} is an indicator function that takes the value 1 if a sanction is imposed and 0 otherwise, and p_{jit} is the sanction that is imposed by individual j on individual i. The payments for all of the rounds will be converted into cash and paid at the end of the experiment.

Note that the parameters were selected such that, without punishment ($p = 0$), freeriding (i.e., pressing the time-out button) is the dominant strategy, as seen in Table 1. An individual who makes an average effort would spend approximately 20 seconds/block on average to produce a correct answer. If the other player does the same, both will earn 10 ECUs. If one presses the time-out button (locks the screen and rests for 20 seconds) while the other works and provides a correct answer, the player who free rides obtains ECUs (and the other receives 5 ECUs). Clearly, 11 ECUs *and* resting is better than 10 ECUs *and* executing the boring task. If both of the players free ride and press the time-out button, they each receive 6 ECUs. However, for a punishment p larger than 1, the only equilibrium of the game is (Count, Count).[8] Because sanctioning is costly, it can be imposed in early rounds by "rational" individuals who play a tit-for-tat strategy, or based on different psychological motives.

When the four minutes have elapsed, the round is over and the participants learn the number of correct answers that they have provided, as well as the number of correct answers that have been provided by their partner. We intentionally choose to disclose only the key performance measure of the partner, which is his or her number of correct answers, and not the number of blocks that were worked on or the number of times that the time-out button was pressed. This information structure of the experiment aims to account for real life situations where partners in team can observe the essentials of the contribution of their team mates to the team's output, but they do not know "everything" about the partner. In particular, it is reasonable to think that partners in a "real" team cannot distinguish between a higher effort and better skills as the explanation for a higher individual output. Of course, the results of the paper depend on the information structure of the experiment, and, if we decided to provide more information to the partners, the results might be different. After the performance is revealed, the player who performed better is asked whether he or she wants to impose the sanction. The other player will then learn the exact amount of the sanction, in addition to the final payoff per se, and a new round starts.

[8] The "maximum punishment" for 20 s would be 30/12 = 2.5.

[Table 1] Payoffs for a typical 20s period (where $p \in [0,2.5]$ is the punishment)

	Count	Rest
Count	10; 10	$(5-0.3p)$; $(11-p)$
Rest	$(11-p)$; $(5-0.3p)$	6; 6

Thus far, the design of the experiment is similar to the "costly sanction" treatment in Vranceanu et al. (2015). However, in this paper, because the focus is on gender interactions, the team members should know the gender of their partner. To convey this information in the least directive way and without relaxing anonymity, we adopt the same method as in Jung and Vranceanu (2017). More specifically, at the very beginning of a given session, the participants were asked to complete an electronic form concerning their "personal characteristics" - age, gender and level of education. Immediately thereafter, the information was communicated to the other team member as "basic information" about his or her partner. In our student population, ages and levels of education are not differentiating characteristics. The only distinctive characteristic was gender. At the end of the experiment, we asked the students whether they could recall the gender of their partner, and if so, what the gender was. A total of 2 out of 132 subjects (one man and one woman) could not recall this.[9]

[Table 2] Sessions and Treatments

Date	Treatment	Number of Subjects	Number of Teams
Oct 23 2014	WW	26	13
Nov 7 2014	MW	24	12
Nov 18 2014	MM	18	9
Dec 5 2014	MW	18	9
Jan 14 2015	MM	14	7
Feb 18 2016	WW	2	1
Feb 18 2016	MW	30	15

To further raise awareness of the gender of the partner, single gender pairs were created in sessions in which only men (only women) were invited; mixed teams were created in mixed-gender sessions. We are aware that single sex environments may make gender stereotypes less salient. This is likely an important weakness of our design.

Table 2 presents the distribution of subjects with respect to sessions and

[9] Some other studies provide the gender information directly. For instance, Sutter et al. (2009) state directly "the subject in the role of A is a woman/man and the subject in the role of B is a woman/man." Other studies indicate the first name of the opponent and let the subject guess the gender. While the latter method might relax anonymity, the former might place too much of an emphasis of the gender issue and thus entail an experimenter demand effect.

treatments: On average, the experiment took approximately 50 minutes. Subjects earned €15.2 on average.

III. Results

3.1. Data, Basic Statistics and Methods

Among the 132 students who participated in the experiment, 64 were female. Thus, 64 individuals were paired with a female partner. Each subject performed the task for 6 rounds; the dataset thus includes 792 observations. Table 2 reports the summary statistics, for the whole sample (first panel) and separately for the population of men and women (second and third panels).

There are two important indicator variables that capture the gender profile of individuals and teams - FE (1 if subject is a woman, 0 if he is a man) and FEp (1 if the partner is a woman, 0 if he is man). NRA is the number of correct answers that were provided by the individual. $DIFSC$ is the difference between the number of correct responses of subject i and that of his or her partner j (i.e., $NRA_i - NRA_j$ in the current round). By the definition of this variable, the full-sample average difference in performance between partners is 0 (a positive performance of individual i with respect to j is offset by the negative performance of j with respect to i). $NBLOCK$ is the number of blocks that were counted by the subject during the round. $NTIMEOUT$ is the number of times that the time-out button was pressed by the subject during the round. $SANC$ is a dummy variable: 1 if the subject penalizes his or her partner conditional on the subject performing better than his or her partner, and 0 if not. $MSANC$ is the amount of penalty that the subject imposes on his or her partner when the former performed better than his or her partner.

To take full advantage of the panel structure of our data and to control for individual characteristics and round specific effects, we estimate the following "generic model"[10]:

$$y_{it} = X_{it}'\beta + \alpha_1 FE + \alpha_2 FEp + \alpha_3 FE \times FEp + u_i + \varepsilon_{it} \qquad (2)$$

where y_{it} stands for the various *outcome variables* pertaining to an individual i

[10] The main reason why we utilize a Random Effects specification rather than a Fixed Effects one stems from our interest in the gender effect, more precisely on the estimates of the gender composition coefficients (FE , FEp , and $FE \times FEp$). Because the gender is fixed for each individual for all of the rounds, in a Fixed Effects specification the individual-specific fixed term would have absorbed gender composition effects.

(at round t), X is the vector of covariates other than the gender indicators. The error term u_i aims to capture the individual random effect and ε_{it} represents the standard error term. Note that in the presence of a significant interaction term $FE \times FEp$, the coefficient α_1 indicates the marginal effect of the subject being a woman, conditional upon FEp being zero (the partner is a man).

[Table 3] Summary Statistics

		Obs	Mean	Sd	Min	Max
FE		792	0.48	0.50	0	1
FEp		792	0.48	0.50	0	1
FExFEp		792	0.21	0.41	0	1
NRA		792	10.10	3.35	0	19
NBLOCK		724	12.16	4.77	1	43
NTIMEOUT		724	0.70	1.69	0	11
DIFSC		792	0.00	3.97	-11	11
SANC		358	0.19	0.39	0	1
MSANC		358	2.64	7.04	0	30
MSANC*cond.*	SANC=1	68	13.91	10.24	1	30
		Male Sample				
FEp		408	0.53	0.50	0	1
NRA		408	10.35	3.44	0	18
NBLOCK		375	13.07	5.76	1	43
NTIMEOUT		375	0.74	1.91	0	11
DIFSC		408	0.42	4.21	-11	11
SANC		207	0.21	0.41	0	1
MSANC		207	3.00	7.69	0	30
MSANC*cond.*	SANC=1	43	14.47	10.99	1	30
		Female Sample				
FEp		384	0.44	0.50	0	1
NRA		384	9.83**	3.22	1	19
NBLOCK		349	11.19***	3.11	2	22
NTIMEOUT		349	0.66	1.41	0	9
DIFSC		384	-0.45***	3.65	-11	10
SANC		151	0.17	0.37	0	1
MSANC		151	2.15	6.01	0	30
MSANC*cond.*	SANC=1	25	12.96	8.94	2	30

Notes: * p<0.10, ** p<0.05, *** p<0.01, t-test comparison.

3.2. The Determinants of Individual Performance

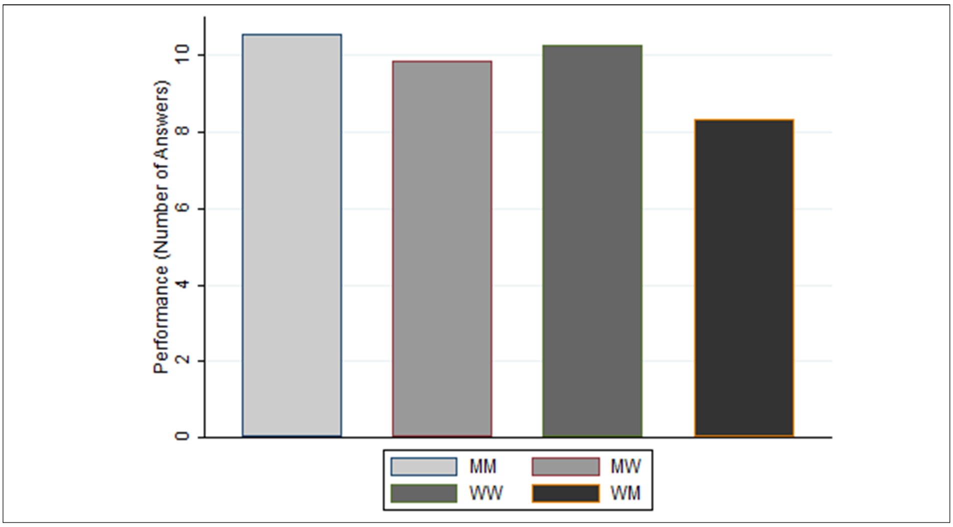
The key measure of performance is the *number of correct answers* that were provided by an individual i in round t (we denoted it by NRA_{it}): The full-sample average performance per round is approximately 10.10 correct answers, with men performing slightly better than women (10.35 vs. 9.83); the difference is nonetheless statistically significant. Our design allows us to move beyond the analysis of gender differences in performance to the analysis of gender interactions - more precisely to the analysis of whether being paired with the same/opposite gender partner has an impact on individual performance. Table 4 and Figure 1 decompose the individual performance (*NRA*) as a total for the whole six rounds of the experiment, in terms of gender compositions: depending on who (a male or a female) was paired with whom (a male or a female).

[Table 4] Individual Performance (Number of Correct Answers)

	Obs	Mean	Sd	Min	Max	MM	MW	WW	WM
MM	192	10.55	3.32	1	18		0.37	0.13	1.17*
						-	(0.39)	(0.06)	(3.35)
MW	216	10.18	3.55	0	18	-0.37		-0.24	0.8
						(0.39)	-	(0.18)	(1.54)
WW	168	10.42	2.55	4	17	-0.13	0.24		1.04*
						(0.06)	(0.18)	-	(2.96)
WM	216	9.38	3.61	1	19	-1.17*	-0.8	-1.04*	
						(3.35)	(1.54)	(2.96)	-

Notes: * p<0.10, ** p<0.05, *** p<0.01.
The stars in the right hand panel indicate whether the mean differs between different gender composition group pairs. The tests are proceeded by regressing the variable of interest on the four group dummies, then use the F test to check for the pairwise difference in coefficients. F values are presented in the parentheses.

[Figure 1] Individual Performance by Gender Composition



These basic statistics suggest that men paired with men[11] have the best individual performance (number of right answers 10.55); when paired with a woman, men's (individual) performance will deteriorate slightly (to 10.18). However, the individual performance of women teamed with women is quite strong (10.42); however, when women are paired with men, their performance is the lowest (9.38).

Regression analysis using a variant of Equation (2) allows us to study the determinants of individual productivity, going beyond these simple descriptive statistics. Table 5 presents the output of the panel data random-effects regression model, with NRA as the dependent variable. As covariates, we include the gender indicator variables, as well as the past period's amount of the sanction ($MSANC_{-1}$) or the sanction dummy variable ($SANC_{-1}$). The number of correct answers in the previous round (NRA_{-1}) is used as a control for the individual's ability in this task, including the net learning effect. Round dummies allow us to capture the residual learning/boredom effect (not captured by the coefficient on NRA_{-1}).

The models in Columns 1 and 2 do not include the past sanction, while the other models do. Because we include one-period lagged variables, the dataset in models 2 to 5 comprises only observations from round 2 to round 6. All of them exclude the two individuals who misspecified their partners.

When the past period performance is not taken into account, the regression in model 1 reveals that the performance of females who are matched with males is the weakest, while it will improve for females who are matched with females, as is indicated by the summary statistics. However, in the regressions where we control for individual ability by introducing past period performance (models 2-5) direct gender differences and gender interactions are not statistically significant. Nevertheless, we can observe that the sanctions that were received in the past have a large impact on current performance. This suggests that gender differences in performance as revealed by the descriptive statistics (Table 2) are explained by different gender-specific responses to past sanctions. Indeed, the analysis of the regression models 2 to 5 reveals the following:

R1. *On average, if a male subject was sanctioned in the previous round, he tends to perform better in the present round.*

This can be inferred from the positive coefficient of both $SANC_{-1}$ or $MSANC_{-1}$, which is statistically significant. Note that in the presence of a significant interaction term $FE \times MSANC_{-1}$, this coefficient on $MSANC_{-1}$ indicates the marginal effect of the past sanction, conditional upon FE being zero

[11] MM: Men paired with men, MW: Men paired with women, WW: Women paired with women, and WM: Women paired with men.

(the subject is a man).

R2. *However, if a woman was sanctioned in the previous round, her performance in the current round will not improve (or it will improve much less).*

This is shown by the marginal effect of the sanction if the subject is a woman, that is the sum of the coefficients of $MSANC_{-1}$ and $FE \times MSANC_{-1}$ in model 4 ($i.e.\ 0.118+(-0.130)\approx 0$); model 3 reveals a modest improvement.

[Table 5] Performance: the random-effects model

	(1)	(2)	(3)	(4)	(5)
FE	-1.209**	-0.355	-0.403*	-0.295	-0.245
	(0.62)	(0.23)	(0.23)	(0.24)	(0.24)
FEp	-0.409	-0.209	-0.205	-0.211	-0.239
	(0.62)	(0.23)	(0.23)	(0.23)	(0.23)
FExFEp	1.487*	0.239	0.267	0.241	0.233
	(0.89)	(0.34)	(0.33)	(0.33)	(0.33)
NRA(-1)		0.776***	0.800***	0.797***	0.790***
		(0.03)	(0.03)	(0.03)	(0.03)
SANC(-1)			0.733**	1.488***	
			(0.31)	(0.49)	
FExSANC(-1)				-1.187*	
				(0.61)	
MSANC(-1)					0.118***
					(0.03)
FExMSANC(-1)					-0.130***
					(0.04)
Round3	1.942***	-0.050	-0.104	-0.090	-0.109
	(0.19)	(0.27)	(0.27)	(0.27)	(0.26)
Round4	2.296***	-0.597**	-0.674**	-0.651**	-0.691**
	(0.19)	(0.27)	(0.27)	(0.27)	(0.27)
Round5	3.635***	0.467*	0.387	0.426	0.425
	(0.19)	(0.28)	(0.28)	(0.28)	(0.28)
Round6	3.542***	-0.664**	-0.731**	-0.687**	-0.696**
	(0.19)	(0.29)	(0.29)	(0.29)	(0.29)
Constant	8.644***	3.443***	3.212***	3.174***	3.279***
	(0.45)	(0.33)	(0.34)	(0.34)	(0.34)
chi2	523.451	890.879	902.965	910.779	920.457
N	780	650	650	650	650
r2_w	0.445	0.153	0.158	0.159	0.160
r2_b	0.036	0.948	0.945	0.946	0.948
r2_o	0.214	0.582	0.585	0.588	0.590

Notes: * p<0.10, ** p<0.05, *** p<0.01.
Individuals who mis-specify their partners' gender are excluded.
We consider only Round2-Round6 in order to take into account the value at $t-1$.

It appears that the motivation crowding-out effect mainly affects female subjects (Rask and Tiefenthaler, 2008).

We also note that conditionally on individual abilities, the round effects are rather mixed.

3.3. The Determinants of Free-Riding Behavior

The analysis of performance can be complemented with an analysis of free-riding behavior, a “choice” that is monitored by the number of times that the subjects press the time-out button (*NTIMEOUT*). With respect to this measure, men tend to take more time-outs (0.74) than women (0.66); however, this difference is statistically insignificant. Nevertheless, does the gender of the partner have an impact on the individual decision to free-ride?

Table 6 and Figure 2 provide the general statistics for (individual) free riding, depending on the gender composition of the teams.

[Table 6] Individual Free-Riding (Number of Time-out Button Pressed)

Obs		Mean	Sd	Min	Max	MM	MW	WW	WM
MM	175	0.46	1.11	0	7	-	-0.52* (2.75)	0.11 (0.46)	-0.44 (2.26)
MW	200	0.98	2.38	0	11	0.52* (2.75)	-	0.63** (4.37)	0.08 (0.04)
WW	154	0.35	0.61	0	3	-0.11 (0.46)	-0.63** (4.37)	-	-0.55* (3.88)
WM	195	0.90	1.77	0	9	0.44 (2.26)	-0.08 (0.04)	0.55* (3.88)	-

Notes: * p<0.10, ** p<0.05, *** p<0.01.
The stars in the right hand panel indicate whether the mean differs between different gender composition group pairs. The tests are proceeded by regressing the variable of interest on the four group dummies, then use the F test to check for the pairwise difference in coefficients. F values are presented in the parentheses.

The comparisons of the average numbers show that when paired with the other gender, subjects tend to free-ride significantly more in comparison to when they are paired with the same gender: men paired with women and women paired with men tend to free-ride more than twice compared to men pairs and women pairs (0.46-0.35 vs. 0.98-0.90).

To control for individual characteristics and round specific effects, we estimate the same type of regression model as in the previous section (equation (2)); however, here we use *NTIMEOUT* as the dependent variable. The sample includes the same 490 observations. Table 7 reports the results of the random-effects panel data regression. The regressions include an interaction term “SAMEGENDER”, which takes a value of 1 if the partners have the same gender (WW or MM) and 0 otherwise.

[Table 7] Determinant of Free-Riding (The Number of Times the Time-Out Button was Pressed): the random-effects model

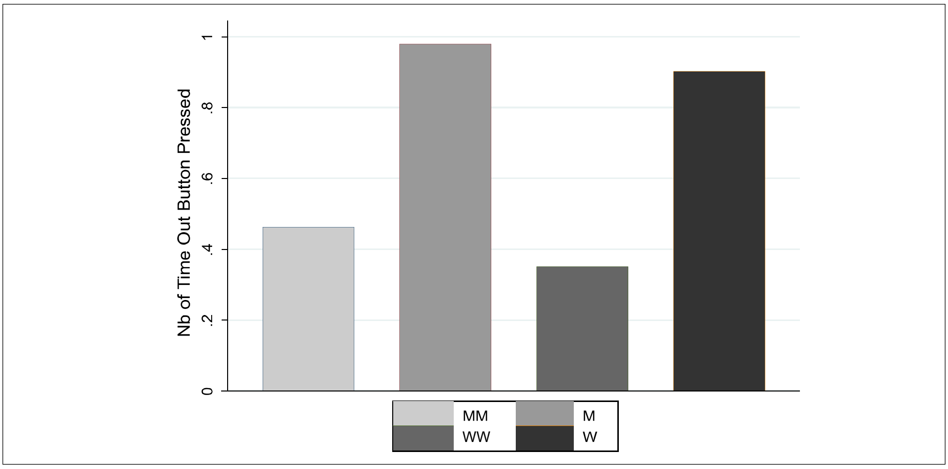
	(1)	(2)	(3)	(4)
	NTIMEOUT	NTIMEOUT	NTIMEOUT	NTIMEOUT
NRA(-1)	-0.110***	-0.110***	-0.122***	-0.121***
	(0.03)	(0.03)	(0.04)	(0.04)
MSANC(-1)	-0.032**	-0.036*	-0.031**	-0.036
	(0.01)	(0.02)	(0.01)	(0.02)
FE	-0.103	-0.110	-0.099	-0.107
	(0.21)	(0.21)	(0.21)	(0.21)
FEp	-0.031	-0.028	-0.035	-0.032
	(0.21)	(0.21)	(0.21)	(0.21)
SAMEGENDER	-0.489**	-0.488**	-0.481**	-0.480**
	(0.21)	(0.21)	(0.21)	(0.21)
FExMSANC(-1)		0.006		0.007
		(0.03)		(0.03)
DIFSC(-1)			0.011	0.010
			(0.03)	(0.03)
FExDIFSC(-1)			0.003	0.006
			(0.04)	(0.04)
Round3	0.176	0.177	0.192	0.192
	(0.16)	(0.16)	(0.16)	(0.16)
Round4	0.492***	0.494***	0.522***	0.523***
	(0.17)	(0.17)	(0.18)	(0.18)
Round5	0.401**	0.400**	0.435**	0.433**
	(0.17)	(0.17)	(0.19)	(0.19)
Round6	0.352*	0.351*	0.403*	0.400*
	(0.19)	(0.19)	(0.22)	(0.22)
Constant	1.823***	1.824***	1.904***	1.905***
	(0.30)	(0.30)	(0.34)	(0.35)
chi2	29.610	29.533	29.821	29.759
N	650	650	650	650
r2_w	0.018	0.019	0.018	0.018
r2_b	0.155	0.155	0.162	0.161
r2_o	0.095	0.094	0.098	0.097

Notes: * p<0.10, ** p<0.05, *** p<0.01.

Individuals who mis-specify their partners' gender are excluded.

We consider only Round2-Round6 in order to take into account the value at $t-1$.

[Figure 2] Individual Free-Riding Behavior by Gender Composition



As expected, better-performing individuals tend to free-ride (press the time-out button) less often. This is indicated by the negative and statistically significant coefficient of NRA_{-1}. Over time, boredom or fatigue causes the participants to press the time-out button more often.

When we turn to gender differences, we note the following:

R3. *Controlling for individual ability, on average the subjects tend to free-ride (press the time-out button) much less when they are paired with the same gender.*

Otherwise, when controlling for past performance and past sanctions, there are no gender specific differences in free riding behavior.

3.4. The Determinants of Sanctioning Behavior

The sanctioning behavior is captured by two measures. From the design section, we know that in any round the best performer in a team can, if he or she wishes, impose a monetary sanction on his or her partner in the range [0;30] ECU, at a cost of 0.30 ECU per ECU of penalty. The sanctions can be applied in 358 cases in which one individual performed strictly better than the other.[12]

As is shown in Table 3, conditional on performing better, 19% of individuals impose penalties on their partners ($SANC = 1$), for an average sanction amount ($MSANC$), conditional on imposing a fine, equal to 13.91 ECU.

Simple comparisons of the data in the second and third panels of Table 3 show

[12] However, we only analyze 353 cases as 5 observations from individuals who misspecify their partner's gender are excluded.

that men tend to apply sanctions more often than women (21% compared to 17%), and they charge on average more than women (14.47 ECUs compared to 12.96 ECUs) conditional on imposing a fine to the partner.

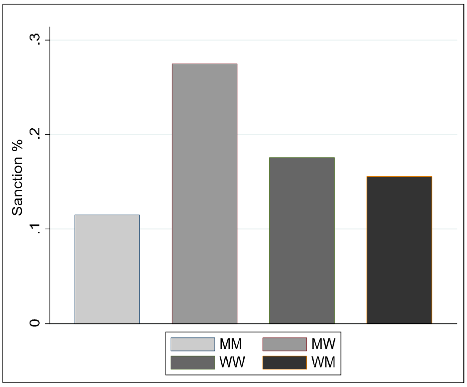
[Table 8] Summary Statistics of Sanction by Gender Composition

	Obs	Mean	Sd	Min	Max	MM	MW	WW	WM
					Sanction Frequency				
MM	87	0.11	0.32	0	1	-	-0.17** (4.20)	-0.07 (0.82)	-0.05 (0.37)
MW	120	0.28	0.45	0	1	0.17** (4.20)	-	0.1 (1.20)	0.12 (1.72)
WW	74	0.18	0.38	0	1	0.07 (0.82)	-0.1 (1.20)	-	0.02 (0.06)
WM	77	0.16	0.37	0	1	0.05 (0.37)	-0.12 (1.72)	-0.02 (0.06)	-
					Amount of Sanction cond. SANC=1				
MM	10	10.80	9.46	2	30	-	-4.78* (3.61)	-1.58 (0.70)	-2.78 (0.76)
MW	33	15.58	11.30	1	30	4.78* (3.61)	-	3.2 (1.40)	2 (1.60)
WW	13	12.38	9.28	2	30	1.58 (0.70)	-3.2 (1.40)	-	-1.2 (0.00)
WM	12	13.58	8.92	5	30	2.78 (0.76)	-2 (1.60)	1.2 (0.00)	-

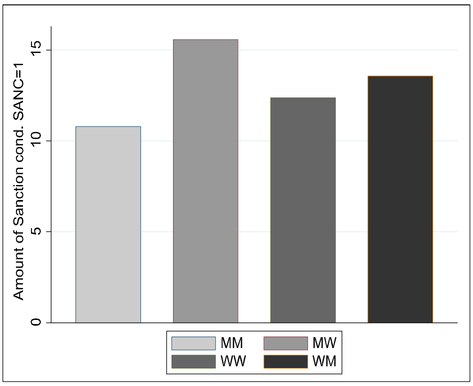
Notes: * p<0.10, ** p<0.05, *** p<0.01.

The stars in the right hand panel indicate whether the mean differs between different gender composition group pairs. The tests are proceeded by regressing the variable of interest on the four group dummies, then use the F test to check for the pairwise difference in coefficients. F values are presented in the parentheses.

[Figure 3] Sanction Frequency by Gender Composition



[Figure 4] Amount of Sanction cond. SANC=1 by Gender Composition



If we now move from gender differences to gender interactions, Table 8, Figures 3 and 4 show the frequency of sanctioning and the average amount of the sanction for each gender type of subject (male, female) depending on the gender of the partner (male, female), for the whole six rounds of the experiment.

The frequency of applying a sanction is the highest in groups in which men perform better than their female partners; the lowest frequency of sanctioning the partner is recorded in all-men teams. In other words, men will, to some extent, "protect" men, and they can be quite harsh with women who underperform.

Additionally, sanctions occur the least when men perform better than their male partners.

On average, the amount of sanctions that are imposed by men on a female partner is much larger than the amount that is imposed on a male partner (15.58 vs. 10.8), and women will impose slightly higher sanctions on men than on women (13.58 vs. 12.38).

As in the previous section, regression analysis can provide additional insights. We estimate several regression models using as a dependent variable either the indicator variable $SANC$ (1 if the subject has applied a sanction) or the ECU amount of the sanction, $MSANC$. There were initially 358 observations for which the sanction option was available (i.e., the subject performed better than his or her partner); however, five observations from the two individuals who misspecified their partners were excluded.

Given that we now control for the difference in performance (recall that on average men perform better in this task), the indicator variables would capture the "plain" gender effect. Table 9 reports the estimation output. We also provide alternative regressions on $SANC$ and $MSANC$ using group gender dummies (MM, MW, WM, WW), which provide additional insights compared to regressions that use individual gender dummies (FE and FEp).

As is expected, the frequency of imposing sanctions and the amount of the sanction is positively related to the difference in the number of correct answers. In other words, the worse the relative performance of the poor performer on the team is, the higher his or her sanction will be (or his or her likelihood of receiving a sanction). This is shown by the positive and significant coefficient on $DIFSC$.

Furthermore,

R4. *When women perform poorly, they tend to be penalized more than men*. This is reected by the positive coefficient of FEp (strongly significant in models 1 and 2).

Instead of using FE and FEp dummies, Models 3 and 4 use indicator variables for types of team. We note the following:

R5. *Men tend to sanction women more often and with a higher amount*. Indeed, taking groups where men are paired with women as the reference, the coefficients for men paired with men and for women paired with men are negative and large. In other words, those who are paired with men sanction significantly less than those who are paired with women.

Rational, self-regarding agents would have no incentive to punish in the last round, as the interaction ends, and the punishment is costly. However, a substantial

[Table 9] Sanction Determinants: the random-effects model

	(1) SANC	(2) MSANC	(3) SANC	(4) MSANC
DIFSC	0.044***	0.907***	0.044***	0.907***
	(0.01)	(0.16)	(0.01)	(0.16)
FE	0.018	-0.078		
	(0.09)	(1.60)		
FEp	0.162*	2.544*		
	(0.08)	(1.51)		
FExFEp	-0.090	-1.374		
	(0.12)	(2.23)		
MW (ref.)				
MM			-0.162*	-2.544*
			(0.08)	(1.51)
WW			-0.072	-1.452
			(0.09)	(1.55)
WM			-0.145*	-2.622*
			(0.08)	(1.54)
Round2	0.089	1.903*	0.089	1.903*
	(0.06)	(0.97)	(0.06)	(0.97)
Round3	0.048	1.453	0.048	1.453
	(0.06)	(0.96)	(0.06)	(0.96)
Round4	0.058	1.447	0.058	1.447
	(0.06)	(0.98)	(0.06)	(0.98)
Round5	-0.081	0.214	-0.081	0.214
	(0.06)	(0.98)	(0.06)	(0.98)
Round6	-0.053	1.479	-0.053	1.479
	(0.06)	(0.98)	(0.06)	(0.98)
cons	-0.026	-2.249*	0.136*	0.295
	(0.08)	(1.34)	(0.07)	(1.29)
chi2	42.164	48.306	42.164	48.306
N	353	353	353	353
r2_w	0.146	0.151	0.146	0.151
r2_b	0.030	0.051	0.030	0.051
r2_o	0.093	0.123	0.093	0.123

Notes: * p<0.10, ** p<0.05, *** p<0.01.

5 observations from Individuals who mis-specify their partners' gender are excluded.

body of literature has shown that “strong reciprocators” will bear the cost of punishment even if they do not expect a positive return only to sanction what they perceive as a deviation from cooperative behavior. Because in our paper sanctions can only be imposed by those who perform better than their partner, we cannot rule out the strong reciprocation assumption. As we can see from Table 10 based only on the last round data, out of 58 occurrences where a player did better than the other, a total of 9 sanctions (15%) were imposed, for an amount of 3.52 ECUs. Even in the last period, men tend to sanction more often and more than women although not to a significant degree. Among male subjects, female partners are sanctioned significantly more often and more severely. Among female subjects, similar patterns are found, although they are not significant.

[Table 10] Descriptive Statistics of Sanction Behavior: Last Round Only

	SANC	MSANC	Obs
Full Sample	0.15	3.52	58
	(0.48)	(1.19)	
Men	0.19	4.16	31
	(0.07)	(1.74)	
Women	0.11	2.78	27
	(0.06)	(1.61)	
Diff	0.08	1.38	
	(0.10)	(2.39)	
MM	0.08	1.54	13
	(0.08)	(1.54)	
MW	0.28	6.06	18
	(0.11)	(2.73)	
Diff	-0.20*	-4.51*	
	(0.15)	(3.48)	
WM	0.07	1.07	14
	(0.07)	(1.05)	
WW	0.15	4.62	13
	(0.10)	(3.12)	
Diff	-0.08	-3.54	
	(0.12)	(3.21)	

Notes: t-test: * p<0.10, ** p<0.05, *** p<0.01.

Table 11 presents the results estimates of the punishment equation, based only on the last round observations.

Most likely, due to the small sample size (58 observations), we do not find any statistically significant gender difference in sanctioning behavior for the last round. However, consistent coefficient signs suggest that, although controlling for the performance, female partners are sanctioned more often and more severely. The last round punishment behavior appears to be consistent with the pattern that is observed in all-round data.

[Table 11] Sanction Determinants: Last Round Only

	(1) SANC	(2) MSANC	(3) SANC	(4) MSANC
DIFSC	0.025	0.969**	0.025	0.969**
	(0.02)	(0.44)	(0.02)	(0.44)
FE	0.017	0.406		
	(0.14)	(3.38)		
FEp	0.204	4.629		
	(0.13)	(3.17)		
FExFEp	-0.109	-0.616		
	(0.19)	(4.62)		
MW(ref.)				
MM			-0.204	-4.629
			(0.13)	(3.17)
WW			-0.092	-0.210
			(0.13)	(3.22)
WM			-0.187	-4.223
			(0.13)	(3.12)
_cons	-0.038	-2.933	0.165	1.696
	(0.13)	(3.16)	(0.12)	(2.86)
N	58	58	58	58
r2	0.091	0.136	0.091	0.136

Notes: * p<0.10, ** p<0.05, *** p<0.01.
5 observations from Individuals who mis-specify their partners' gender are excluded.

3.5. The Analysis of Stereotypes

As noted above, at the end of the experimental session the participants were asked to answer three questions regarding their gender preference and use a five-item scale to state their opinions (from very poor (1) to very good (5)). The sample includes the 98 participants who were aware of the gender of their partner. Table 12 presents the regression models using these expressed preferences as a dependent variable. The $AvDIFSC$ is the all-round average difference in an individual's score (positive if he or she outperformed, negative if he or she under-performed), and $FEp \times AvDIFSC$ is the interaction term with the partner being female.

The first column presents the results for the question "How do you evaluate your partner's performance?" While many factors are insignificant, we find the following:

R6. *The better the relative performance of an individual, the lower the assessment of his or her partner's performance (and vice-versa).*

In line with intuition, the negative coefficient of $AvDIFSC$ corroborates the consistency of the subjective scale with the objective performance of the task.

[Table 12] Preference Questionnaire about Partner/Gender

	(1)	(2)	(3)
	Performance	Performance Other	Punishment Other
	Partner	Gender	Gender
FE	-0.227	0.042	-0.233
	(0.23)	(0.21)	(0.30)
FEp	-0.283	0.220	0.440**
	(0.26)	(0.19)	(0.20)
FExFEp	0.488	-0.261	0.493
	(0.34)	(0.25)	(0.33)
FEpxAvDIFSC	-0.109	-0.038	-0.139***
	(0.07)	(0.05)	(0.05)
AvDIFSC	-0.181***	0.059	0.056*
	(0.05)	(0.04)	(0.03)
_cons	3.906***	3.000***	3.031***
	(0.14)	(0.15)	(0.15)
N	98	98	98
r2	0.474	0.083	0.245

Notes: * p<0.10, ** p<0.05, *** p<0.01.
Bootstrapping 1000 replications.
2 Individuals who mis-specify their partners' gender are excluded.
avDIFSC is the average of DIFSC for each individual.

The negative coefficient of the interaction term $FEp \times AvDIFSC$ would suggest that, all things being equal, subjects tend to be more critical of women than they are of men by under-estimating the ex-post performance of women.

The second column pertains to the question "In this experiment, considering the performance of your partner, do you believe that a [opposite gender as partner] would have performed" for which we did not observe any significant results. Based solely on the sign of coefficient of FEp, if the subject was paired with a woman, she or he tended to believe that if she or he had been paired with a man, the male partner would have performed better (and this belief is stronger for men, as is shown by the sign of $FE \times FEp$). This suggests that subjects naturally underestimate the performance of women.

The final column reports the results for the question "If your partner had been of the opposite gender, how do you imagine she or he would have penalized you?"

R7. *As is shown by the positive and significant coefficient of* FEp*, in the case of males who are paired with females, they believe that if the partner had been a man, he would have imposed heavier sanctions.*

Subjects tend to believe that women would penalize less than men, which is in line with the observed behavior in this experiment.

These answers suggest that negative gender stereotypes in the workplace still exist, and they affect not only the men but also the women themselves.

IV. Conclusion

This paper reports results from a team production, real-effort experiment in which the gender of the partners is common knowledge. The purpose of this design is to analyze the role of gender interaction on individual performance and sanctioning behavior. Partners on a team receive an equal share of the team's output as compensation. The best performer on a team is given the option to impose a monetary sanction on the less-effective partner. Teams differ in their gender composition. We can therefore observe individual behavior in all-men, all-women and mixed teams.

Among the main results, we emphasize the following:

(1) When they underperform, all other things being equal, women are sanctioned more often and more heavily than men;

(2) If they are sanctioned, women's performance will not improve, whereas men tend to improve their performance;

(3) Free riding is more acute in mixed-gender teams compared to same-gender teams.

These differences in sanctioning behavior (women are sanctioned more heavily) combined with gender-different responses to sanction would explain gender differences in individual performance in a repeated interaction. We observed that men tended to improve their performance from one round to another when they were sanctioned, and the gain in performance was positively related to the amount of the sanction. Thus, in all-men teams, internal punishment should contribute to an improvement in the team's performance. As the sanction is proportional to the difference in scores, in the long run, the gap in the performance of the two team members should narrow to the lowest natural difference in abilities, with sanctions gradually declining and stabilizing. However, we have shown that women do not respond to sanctions by improving their performance (it may even slightly decrease). Furthermore, should they under-perform, they receive larger sanctions than men, and these sanctions are the most substantial when women are paired with men. Thus, in mixed teams where a man outperforms his female partner, the punishment might be counterproductive by deteriorating motivation and reducing effort in a repeated interaction.

Compared to other real-effort tasks that have been used in team production experiments (such as coding, adding numbers, and solving mazes), this task (counting 7s) does not require any particular reasoning skills; however, it draws heavily on attention, focus and resistance to boredom. One might argue that according to stereotypes, "counting" is a task where men are expected to do better than women, which would bias the results. While it seems difficult to sustain such a belief in our population of students who were recruited through a tough national competitive exam (including math), it would be interesting in future research to study whether the main results hold if the participants are required to perform a different task. Additionally, our results, which were observed in two-person teams, might be less prominent in larger groups, in particular because the incentive to free-ride on the punishment would be stronger.

As with all experiments, extreme caution is required with attempts to extrapolate conclusions from such a simple experiment to real life situations. With this caveat, our analysis can shed some light on the dynamics of performance in teams. There are many sectors and enterprises in which production has historically been primarily completed by men. This is true not only for many mass-production processes in manufacturing and construction but also in many clerical jobs (consulting, law firms, and even in academia). It is easy to understand why a norm of punishing defectors should emerge in such work environments. With the accelerated feminization of many organizations (Bowles et al., 2007), the gender composition of teams is changing rapidly. However, if norms do not change at the same pace, and workers uniformly impose internal punishment, firm performance may well stagnate. Given the changing work environment conditions, it might a good time to reconsider the value of forgiveness.

Acknowledgments This project was carried out within the research framework of the Labex MME-DII (http://labex-mme-dii.u-cergy.fr). The authors would like to thank ESSEC Research Center and ESSEC Alumni Foundation for having funded this research and Delphine Dubart from the ESSEC Experimental Lab for her technical assistance. They would like to thank two anonymous referees, Maia Linask, Estefania Santacreu Vasut, Agnieszka Wysokinska, as well as participants to the *ESSEC Behavioral Workshop* in Cergy in March in 2015, to the *THEMA Research Seminar* in Cergy in April in 2015, to the *Annual Conference of the French Association of Experimental Economics (ASFEE)*, in Paris on June 14-15 in 2015 and to the *Research Workshop of the GATE* in Lyon, on June 19th in 2015 for their suggestions and remarks on an early draft.

Appendix

Instructions[13]

Slide 1.

Good morning. Thank you for participating in this experiment. Please read these instructions carefully and, should you have any questions, raise your hand and call the administrator. Communication between participants is forbidden. Please turn off cellular phones. A payment in cash will be provided at the end of the session.

Slide 2. Personal characteristics

-You are: [A man / A woman]
-Your age is [X]
-Your education level is: [Baccalaureate+1,+2,+3,+4+5,+6 or more years of education]

Slide 3. About your partner

In this experiment you will be paired at random with another person present in this room; this pairing is strictly anonymous.
-The gender of the partner is: [man/woman]
-The age of the partner is: [X]
-The education level of the partner is: [Baccalaureate+1,+2,+3,+4+5,+6 or more years of education]

Slide 4. Main rule

-You will be required to perform an effort task jointly with a partner, during 6 identical rounds of the same experiment.
-Each round lasts for 4 minutes; the clock starts when you open the first active screen, and stops after 4 minutes. During a round, the remaining time is displayed in red characters, in the upper right corner of the screen (in seconds).
-A payoff in euros will be delivered at the end of the experiment. The payment is connected to performance in the task, according to a rule known to everyone.
-Partners will be matched in pairs at random. Your partner will not change from one round to another. His identity will not be revealed to you.
-During each round the computer displays a sequence of blocks of figures (0 to 9) in six lines and 30 columns. Your task is to accurately count how many times the figure 7 appears in a block and then report this number in a box. The answer is considered correct if it corresponds to the right number of 7s in the block, with a tolerated error margin of 1. For instance, if the correct number of 7s is 30, answers

[13] Translated from French.

29, 30 and 31 will be considered correct.

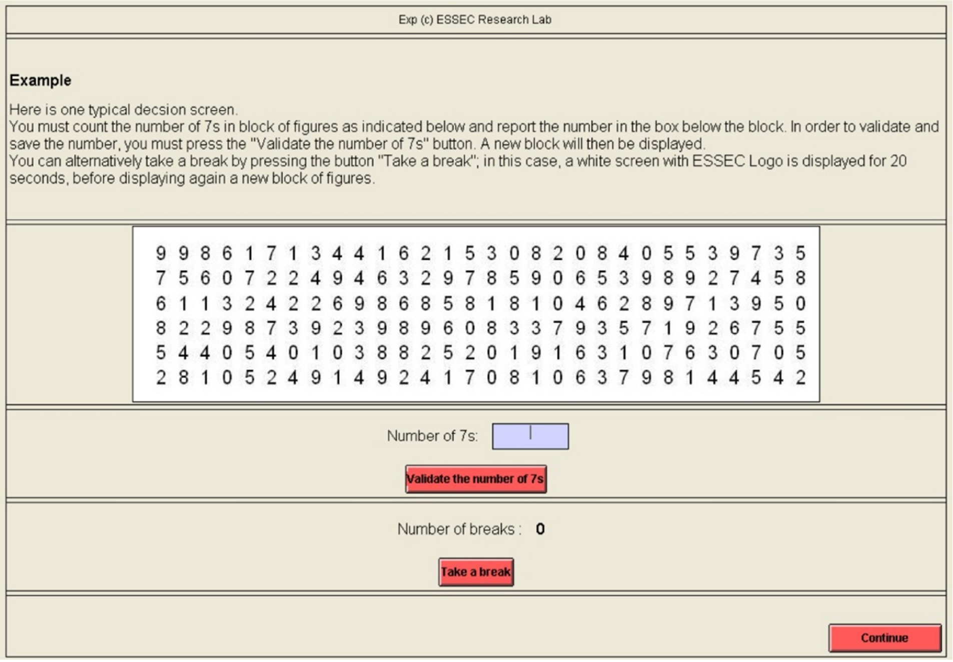
-Once the counted number of 7s is recorded in the box, you must press the “validate” button, to save it. After you click, a new random block of figures is automatically generated and the effort task can continue.

-At any moment you can take a break by pressing the button “Take a break”. The break stops the counting task for 20 seconds; a screen with ESSEC logo appears. If the round stops in less than 20 seconds, breaks are no longer possible.

-At the end of each round the computer will display the total number of correct answers that you have provided and the total number of correct answers provided by your partner.

-At the end of each round, before moving to the next round, the player who provided the highest number of correct answers can, if he/she wants so, impose a fine on his/her partner. The decision belongs to him/her, it is not compulsory to impose the fine. In the event that players have provided an identical number of correct answers, no sanction is possible.

Slide 5. The example slide - main decision screen



Exp (c) ESSEC Research Lab

Example

Here is one typical decsion screen.
You must count the number of 7s in block of figures as indicated below and report the number in the box below the block. In order to validate and save the number, you must press the "Validate the number of 7s" button. A new block will then be displayed.
You can alternatively take a break by pressing the button "Take a break"; in this case, a white screen with ESSEC Logo is displayed for 20 seconds, before displaying again a new block of figures.

```
9 9 8 6 1 7 1 3 4 4 1 6 2 1 5 3 0 8 2 0 8 4 0 5 5 3 9 7 3 5
7 5 6 0 7 2 2 4 9 4 6 3 2 9 7 8 5 9 0 6 5 3 9 8 9 2 7 4 5 8
6 1 1 3 2 4 2 2 6 9 8 6 8 5 8 1 8 1 0 4 6 2 8 9 7 1 3 9 5 0
8 2 2 9 8 7 3 9 2 3 9 8 9 6 0 8 3 3 7 9 3 5 7 1 9 2 6 7 5 5
5 4 4 0 5 4 0 1 0 3 8 8 2 5 2 0 1 9 1 6 3 1 0 7 6 3 0 7 0 5
2 8 1 0 5 2 4 9 1 4 9 2 4 1 7 0 8 1 0 6 3 7 9 8 1 4 4 5 4 2
```

Number of 7s:

Number of breaks : **0**

Slide 6. Compensation rule

Gains are denominated in Experimental Currency Units (ECU). For each round, the payoff for one player is made up of three elements:

1. The compensation related to the effort task
2. A gain provided when taking a break
3. Less the penalty (if any)

Which are:

1. For each player, the ECU compensation related to the effort task is equal to half the total number of correct answers provided by the team during the round, times 10. For instance, if player 1 provided 8 correct answers and player 2 provided 5 correct answers, the gain for each player related to the effort task is 0.5x(8+5)x10=65 ECUs
2. For each 20 second break, you will receive 6 ECU, whatever your compensation for the effort task.
3. At the end of each round, before the next round starts, the player who provided the largest number of correct answers can, if he wants to impose a fine on his partner, for an amount between 1 and 30 ECUs. The gain of the partner is reduced by that amount. One ECU in fines will cost the punisher 0.30 ECU. (No sanction is possible if players provide the same number of correct answers)

-At the end of the experiment, the total amount in ECU will be converted into euros at the exchange rate 100 ECU=2.5 euros.

Slide 8. Check questions

To make sure that you have understood well the rules of the game, please answer these questions:

Case 1.

During the round you got 4 right answers and your partner got 2 right answers. You took two breaks.

Your gain in ECU is:

(a) 0.5*(4+2)*10+2*6; (b) (4+2)*10 + 2*6; (c) I do not know

Case 2.

At the end of the round you got 4 right answers and your partner got 2 right answers.

Can you impose a fine on your partner?

(a) Yes (b) No (c) Don't know

Can your partner impose a fine on you?

(a) Yes; (b) No; (c) Don't know

Case 3.

At the end of the round you chose to impose a fine on your partner. The amount of the fine can be:

(a) Between 1 and 10; (b) Between 1 and 30; (c) I do not know.

Slide 9.
Correct answers are:

Case 1.
-During the round you got 4 right answers and your partner got 2 right answers. You took two breaks.
-The right answer is: you have half of the total points times 10, and the compensation for the breaks (2x6 ECUs), that is a total of 0.5(4+2)*10+2x6

Case 2.
-You got 4 right answers and your partner got 2 right answers.
-Yes, you can impose a fine on him. Attention, this is an option; you do not need to impose a fine.
-No, he cannot impose a fine on you

Case 3.
At the end of the round you chose to impose a fine on your partner. The amount of the fine can be between 1 and 30. The payoff of your partner will be reduced by this amount.

Slide 10.
-If you have any questions, please raise your hand and address it to the administrator.
-If you are sure you have understood the rules of the game, you can press the button below to launch the experiment.
-The experiment starts when all subjects have pressed the button.

Slide 11. (Main decision screen)
Similar to "Example" in Slide 5 (but without the text on top of the screen; and the timer on the right upper corner).

Slide 12. Results on task
-Your performance: Number of counted blocks [ ], Number of right answers [NRA1], Number of breaks [ ]
-The performance of your partner: Number of right answers [NRA2].
-[If NRA1 > NRA2 the computer displays] Do you want to impose a fine on the partner?
-You choose: Yes // No
-If you click Yes the computer displays "choose the amount of the fine" [A=1 to 30], then "Validate"

Slide 13. Payoff for the round

-Your partner has imposed a fine on you [or /did not impose a fine on you]
-The amount of the fine is: [A]
-Your payoff for the round is: [ ]

At the end of the experiment (after the 6 rounds)

Question 1.

Please evaluate the performance of your partner:
-[very poor, poor, average, good, very good]

Question 2.

Your partner was:
-[A man, A woman, Don't know]

Question 3.

In this experiment, considering the performance of you partner, you believe that a [opposite gender partner] would have performed:
-[much worse, worse, the same, better, much better]

Question 4.

In this experiment, do you think that a [opposite gender partner] would have applied a sanction:
-[much lower, lower, identical, higher, much higher]

Last slide

Thank you for having participated in this experiment.
The total gain for the experiment is [ ] euros.

References

Alchian, A. A. and H. Demsetz (1972), "Production, Information Costs, and Economic Organization," *American Economic Review*, 62(5), 777-795.

Apesteguia, J., G. Azmat, and N. Iriberri (2012), "The Impact of Gender Composition on Team Performance and Decision Making: Evidence from the Field," *Management Science*, 58(1), 78-93.

Asiedu, E. and M. Ibanez (2014), "The Weaker Sex?: Gender Differences in Punishment Across Matrilineal and Patriarchal Socienties," *Leibniz-Informationszentrum Wirtschaft, GlobalFood Discussion Papers Nr 30*.

Barron, J. M., Gjerde, and K. Paulson (1997), "Peer Pressure in an Agency Relationship," *Journal of Labor Economics*, 15(2), 234-254.

Ben-Ner, A., F. Kong, and L. Putterman (2004), "Share and Share Alike? Gender-pairing, Personality, and Cognitive Ability as Determinants of Giving," *Journal of Economic Psychology*, 25(5), 581-589.

Boning, B., C. Ichniowski, and K. Shaw (2007), "Opportunity Counts: Teams and the Effectiveness of Production Incentives," *Journal of Labor Economics*, 25(4), 613-50.

Bowles, H. R., L. Babcock, and L. Lai (2007), "Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes it Does Hurt to Ask," *Organizational Behavior and Human Decision Processes*, 103(1), 84-103.

Carpenter, J. P. (2007), "Punishing Free-riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods," *Games and Economic Behavior*, 60(1), 31-51.

Dickinson, D. and M.-C. Villeval (2008), "Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-out Theories," *Games and Economic Behavrior*, 63(1), 56-76.

Dittrich, M., A. Knabe, and K. Leipold (2014), "Gender Differences in Experimental Wage Negotiations," *Economic Inquiry*, 52(2), 862-873.

Dufwenberg, M. and A. Muren (2006), "Generosity, Anonymity, Gender," *Journal of Economic Behavior and Organization*, 61(1), 42-49.

Eckel, C. C. and P. J. Grossman (1998), "Are Women Less Selfish than Men?: Evidence from Dictator Experiments," *Economic Journal*, 108(448), 726-735.

Eckel, C. C. and P. J. Grossman (2001), "Chivalry and Solidarity in Ultimatum Games," *Economic Inquiry*, 39, 171-188.

Falk, Armin, E. F. and U. Fischbacher (2005), "Driving Forces Behind Informal Sanctions," *Econometrica*, 73(6), 2017-2030.

Fehr, E. and S. Gächter (2000), "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), 980-994.

Fehr, E. and J. List (2004), "The Hidden Costs and Returns of Incentives - trust and Trustworthiness Among CEOs," *Journal of the European Econonmic Association*, 2, 743-771.

Fehr, E., M. Naef, and K. M. Schmidt (2006), "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment," *American*

Economic Review, 1912-1917.

Fehr, E. and B. Rockenbach (2003), "Detrimental Effects of Sanctions on Human Altruism," *Nature*, 422, 137-140.

Fershtman, C. and U. Gneezy (2001), "Discrimination In A Segmented Society: An Experimental Approach," *The Quarterly Journal of Economics*, 116(1), 351-377.

Festré, A. and P. Garrouste (2014), "Theory and Evidence in Psychology and Economics about Motivation Crowding Out: A Possible Convergence?," *Journal of Economic Surveys, On-line in-print, http://dx.doi.org/10.1111/joes*, 12059.

Fischbacher, U. (2007), "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171-178.

Frey, B. S. and I. Bohnet (1995), "Institutions Affect Fairness: Experimental Investigations," *Journal of Institutional and Theoretical Economics*, 286-303.

Frey, B. S. and R. Jegen (2001), "Motivation Crowding Theory," *Journal of Economic Surveys*, 15(5), 589-611.

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

Hederos Eriksson, K. and A. Sandberg (2012), "Gender Differences in Initiation of Negotiation: Does the Gender of the Negotiation Counterpart Matter?," *Negotiation Journal*, 28(4), 407-428.

Holmstrom, B. (1982), "Moral Hazard in Teams," *Bell Journal of Economics*, 13(2), 324-340.

Hoogendoorn, S., H. Oosterbeek, and M. V. Praag (2013), "The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment," *Management Science*, 59(7), 1514-1528.

Houser, D., E. Xiao, K. McCabe, and V. Smith (2008), "When Punishment Fails: Research on Sanctions, Intentions and Non-cooperation," *Games and Economic Behavior*, 62(2), 509-532.

Itoh, H. (1991), "Incentives to Help in Multi-agent Situations," *Econometrica*, 59(3), 611-36.

Ivanova-Stenzel, R. and D. Kübler (2011), "Gender Differences in Team Work and Team Competition," *Journal of Economic Psychology*, 32(5), 797-808.

Jung, S. and R. Vranceanu (2017), "Experimental Evidence on Gender Interaction in Lying Behavior," *Revue Economique, Forthcoming*.

Kandel, E. and E. P. Lazear (1992), "Peer Pressure and Partnerships," *Journal of Political Economy*, 100(4), 801-817.

Kirstein, A. (2008), "Bonus and Malus in Principal Agent Relations with Fixed Pay and Real Effort," *Fixed Wage*, 60, 280-303.

Kuhn, P. and M.-C. Villeval (2015), "Are Women More Attracted to Co-operation than Men?," *Economic Journal*, 125, 115-140.

Lamiraud, K. and R. Vranceanu (2015), "Group Gender Composition, Tolerance to Risk and Economic Performance: New Evidence from an Original Business Game," *ESSEC Working Paper*, 1515 .

Lawler, E. E. I. and S. A. Mohrman (2003), "Pay Practices in Fortune 1000 Corporations," *Center for Effective Organizations, Publication G*, 448, 3-20.

Legros, P. and S. A. Matthews (1993), “Efficient and Nearly-efficient Partnerships,” *Review of Economic Studies*, 60(3), 599-611.

McAfee, R. P. and J. McMillan (1991), “Optimal Contracts for Teams,” *International Economic Review*, 32(3), 561-77.

Mohnen, A., K. Pokorny, and D. Sliwka (2008), “Transparency, Inequity Aversion, and the Dynamics of Peer Pressure in Teams: Theory and Evidence,” *Journal of Labor Economics*, 26(4), 693-720.

Niederle, M. and L. Vesterlund (2007), “Do Women Shy Away from Competition? Do Men Compete too Much?,” *Quarterly Journal of Economics*, 122(3), 1067-1101.

Niederle, M. and L. Vesterlund (2011), “Gender and Competition,” *Annual Review in Economics*, 3, 601-630.

Nikiforakis, N. and H.-T. Normann (2008), “A Comparative Statics Analysis of Punishment in Public-good Experiments,” *Experimental Economics*, 11(4), 358-369.

Pokorny, K. (2008), “Pay-but do not Pay too Much: An Experimental Study on the Impact of Incentives,” *Journal of Economic Behavior and Organization*, 66(2), 251-264.

Rask, K. and J. Tiefenthaler (2008), “The Role of Grade Sensitivity in Explaining the Gender Imbalance in Undergraduate Economics,” *Economics of Education Review*, 27(6), 676-687.

Solnick, S. J. (2001), “Gender Differences in the Ultimatum Game,” *Economic Inquiry*, 39(2), 189-200.

Sutter, M., R. Bosman, M. Kocher, and F. Winden (2009), “Gender Pairing and Bargaining? Beware the Same Sex!,” *Experimental Economics*, 12(3), 318-331.

Vranceanu, R., F. E. Ouardighi, and D. Dubart (2015), “Team Production with Punishment Option: Insights from a Real-effort Experiment,” *Managerial and Decision Economics*, 36(6), 408-420.