Sabotaging a Teammate without Reason: Experiments on the Impact of Bonus Payments

Sookie Xue Zhang* Ralph-Christopher Bayer†

January 30, 2025

Abstract

We conduct laboratory experiments to examine how sabotage behaviour in teams evolves in equivalent environments with and without bonus payments. We use a symmetric two-player coordination game where deviating from the efficient equilibrium represents sabotage behaviour, as it leaves a player's payoff unchanged but imposes positive costs for the other player. In our experimental treatments, the sabotage payoff is framed as resulting from a bonus, while in the control treatment, no bonus is involved. In a further treatment variation, we either frame the situation as a game with a rival or with a teammate. We find that introducing a bonus (without changing material incentives) and a rival framing both lead to more sabotage behaviour and harm efficiency. Our results suggest that bonuses can lead to more harmful sabotage behaviour even if they do not provide material incentives for it.

Keywords: Bonus Payments; Framing; Social Preferences; Experiments

*Corresponding author. Wenlan School of Business, Zhongnan University of Economics and Law, Wuhan 430064, China. E-mail address: sookie.xue.zhang@outlook.com.

†School of Economics and Public Policy, The University of Adelaide, SA 5005, Australia. E-mail address: ralph.bayer@adelaide.edu.au.

1 Introduction

Bonuses, awards, and relative pay are widely used to financially incentivise individuals to exert efforts in teams (Lazear and Rosen, 1981). In addition to the positive incentives for effort, they can also have unintended and less desirable consequences. Recent experimental studies indicate that bonus-based incentive schemes can backfire in workplaces by increasing cheating (Gill et al., 2013), absenteeism (Alfitian et al., 2023), and crowding out intrinsic motivation through more cautious goal setting (Gonzalez et al., 2020). When bonuses or prizes are awarded based on relative rankings, destructive behaviour against competitors, i.e. sabotage, is often a rational choice (Lazear, 1989; Chen, 2003; Grund and Sliwka, 2005). Laboratory experiments (Harbring and Irlenbusch, 2005, 2008, 2011) and field experiments (Carpenter et al., 2010; Flory et al., 2016) have shown that the extent of sabotage increases with bonuses and relative payment schemes that provide sabotage incentives. For a more detailed survey, see Gangadharan et al. (2020). Moreover, Charness et al. (2014) show that sabotage activity increases in a flat-wage task if feedback on relative performance ranking is given to participants[1]

This paper is motivated by the conjecture that prizes and bonuses could impact sabotage behaviour that goes beyond that generated by financial incentives. Buser and Dreber (2016) show that a winner-takes-it-all real-effort contest harms cooperation in a subsequent public goods game. Similarly, Burks et al. (2009) show in the field that bicycle couriers are less prosocial to other bicycle couriers if they are paid according to performance despite the fact that performance pay in this context does not provide material incentives to behave less prosocially. Studies in organizational psychology using experiments show that pay-for-performance activates competitive norms. Correlational field studies corroborate the results Gläser et al. (2017, 2022, e.g.).[2] From this we conclude that it is possible that the sheer existence of prizes and bonuses yields an extra incentive to sabotage others in a team environment through priming and the activation of more competitive preferences. For this, it is not necessary that the bonuses actually provide material sabotage incentives. Hence, under this premise, theories that predict that sabotage in teams is solely due to incentives underestimate the dark side of competition induced by bonuses, prizes, and relative pay.

We design a novel sabotage game to test the hypothesis that bonuses and prizes lead

[1]Observational field data on sabotage behaviour are rare. Exceptions are Drago and Garvey (1998) and Deutscher et al. (2013).

[2]Experimental studies in organizational psychology typically do not satisfy the classical methodological rules in experimental economics as they are often not incentivised or use deception.

to more efficiency-damaging behaviour even if it does not provide material incentives to sabotage. In our sabotage game, two players simultaneously chose to sabotage each other or not. Sabotage reduces the payoff of the other player, leaves the own payoff unchanged if the other player does not sabotage but marginally reduces the damage of the other player sabotaging. This yields an interesting game with two Nash Equilibria: both players refraining from sabotaging and both players sabotaging. The non-sabotage equilibrium is in weakly dominated strategies but payoff-dominant, while the sabotage equilibrium is weak and risk-dominant. In this game competitive preferences and preferences for social efficiency or altruism select a different equilibrium by rendering sabotage or not to sabotage strictly dominant. The slightest social preferences with a warm glow for the other player or for social efficiency select the non-sabotage equilibrium, while competitive preferences make the sabotage equilibrium unique. This game is an ideal test-bed for our conjecture that bonuses activate more competitive preferences, which should lead to more sabotage behaviour in our game.

In our control treatment, the baseline game is played under a team frame. The other player is referred to as a "teammate." In our first experimental treatment (i.e. the *bonus* treatment), we introduce a bonus without changing the payment of the game or the framing. Now sabotaging, when the other player does not sabotage, yields a sizeable bonus for the saboteur.[3] In order to keep the payoff structure identical to the baseline, unilateral sabotage incurs an offsetting cost, which is identical to the bonus. In our second experimental treatment (i.e. the *competitive bonus* treatment), the bonus treatment is accompanied by a competitive framing. The other player is called "rival" and the bonus is explained in terms of competition: " the player who has a higher payoff than the rival wins an extra bonus." The game remains identical in terms of the final payoff to those played in the two other treatments.

The results confirm our hypothesis. Despite the fact that participants play the same game in all three treatments, behaviour differs. In particular, the level of surplus is lower in the bonus treatment than in the control treatment and lower again in the competitive bonus treatment. This is mainly driven by different relative frequencies with which the efficient non-sabotage equilibrium and the inefficient sabotage equilibrium are played. Disequilibrium play occurs at similar rates in the three treatments. We conclude from our findings that bonuses can lead to increased sabotage in teams even if they do not provide material

[3]Actions are denoted neutrally as A and B.

incentives for it. This effect is particularly strong if the bonuses are combined with a competitive framing. Our dynamic analysis shows that these differences are persistent rather than transitory. There are important policy implications for designers of remuneration systems and competitions. It is not sufficient to make sure that a system does not provide material incentives for sabotage. An additional check is required if a remuneration system activates competitive preferences, which might lead to damaging sabotage behaviour. Our experiments show that this can easily happen when the system contains bonuses.

The remainder of the paper is as follows. The next Section explains the underlying game and the resulting design in detail and develops our hypotheses. Section 3 presents our results and the final Section offers a short conclusion.

2 Experimental Design

We intend to investigate whether the existence of bonuses increases instances of sabotage, even if they do not provide a financial incentive to do so. For this, we use a simple sabotage game as a starting point.

2.1 The game

Two players each have an endowment of e. They decide simultaneously to perform a sabotage action (S) or not (N). Sabotaging the other player is costless for them, but reduces the other player's payoff by $s < e$. The resulting game is strategically uninteresting since a player's choice of action does not impact her own payoff. This implies that any admissible strategy profile (pure or mixed) constitutes a Nash equilibrium. In order to obtain richer predictions for the baseline game, we add a small amount $\epsilon < s/2$ to both payoffs if both players are sabotaging. This can be interpreted as sabotaging marginally reduces the harm of being the recipient of sabotage. This transforms the sabotage game into a "Harm-Thy-Neighbor" game. Table 1 represents the normal form of the resulting game.

	S	N
S	$e-s+\epsilon, e-s+\epsilon$	$e, e-s$
N	$e-s, e$	e, e

Table 1: The baseline game

Both players choosing not to sabotage Pareto-dominates all other outcomes. The surplus decreases with the number of sabotaging actions taken. The two Nash equilibria of this game are (N, N) and (S, S), resulting in the most efficient and least efficient outcome, respectively.[4] The very modest payoff differences for a player caused by choosing S or N, i.e. ϵ in the case the other player chooses S, allow for plenty of room for factors beyond material incentives (like social preferences and framing) to shape choices.

Table 2 shows the actual game we used. For the experiments, we neutrally labeled the actions as A (sabotage) and B (not sabotage). Moreover, we chose $e = 98$ and $s = 40$, which makes sabotage very damaging relative to the endowment. Focal points due to prominent numbers are avoided. Finally, we chose $\epsilon = 1$, so that deviating from the sabotage equilibrium is not particularly costly.

	A	B
A	59 59	98 58
B	58 98	98 98

Table 2: The payoff structure

Before we discuss the different treatments, we will briefly comment on how the choice of game relates to the typically expected play in a framing-free version of this game. In coordination games, there are two main (often competing) determinants of which equilibrium is typically played. Both were introduced in Harsanyi et al. (1988). The first criterion is how risky a strategy is in terms of the damage incurred by a player who plays the equilibrium strategy, while the other player only plays the equilbrium strategy with probability half. The less risky equilibrium is called risk-dominant. On the other hand, an equilibrium might be more attractive if all players receive higher payoffs than in the equilibrium (payoff dominance). When the two criteria yield conflicting predictions, it is a priori unclear which criterion should dominate. Harsanyi and Selten theorized that payoff-dominance should be the overriding criterion, while evolutionary approaches such as Kandori et al. (1993) favour the risk-dominant equilibrium. Empirically, both criteria matter and the payoff dominant equilibrium is played more often if it is also risk-dominant (Dal Bó et al., 2021). On the other hand, Battalio et al. (2001) show that the risk-dominant equilibrium is played less often if the optimization incentives are low (i.e. the cost of failing to best respond to the equilibrium play of the other player is low). Similarly, the closer the riskiness of the risk-

[4]The weak dominance of S rules out the existence of a mixed strategy equilibrium.

dominant strategy is to that of the payoff-dominant strategy, the less likely it is that the risk-dominant equilibrium is played (Dubois et al., 2012).

In our game, the sabotage equilibrium (AA) is risk dominant, while the non-sabotage equilibrium (BB) is payoff dominant. The risk-dominance of (AA) is very weak and the optimization incentives are low. In contrast, the pay-off dominance is strong. Hence, our game in its basic form is expected to yield a large fraction of non-sabotage choices of B. However, given that the sabotage equilibrium is risk-dominant, we would expect at least some sabotage (A). Moreover, we would expect a slight increase in the fraction of sabotage actions, since evolutionary and learning dynamics favour the risk-dominant equilibrium. Behaviour like that provides a nice benchmark that leaves sufficient room for framing to shift behaviour to or away from sabotage.

Finally, our base game is designed to emphasize one specific channel through which framing changes play. Framing might change a player's belief about the likely play of the other player or it might activate a different set of social preferences. Given the very small payoff differences between playing A and B for a player conditional on the action of the other player, the expected payoff difference between the two strategies varies minimally (between 0 and 1 experimental currency units) with the beliefs. Hence, the impact of beliefs on the relative expected payoffs from playing A or B is negligible. For this reason, our design limits the channel through which sabotage frequencies are affected by changes in preferences.

Slight changes in social preferences radically change the predictions. Any kind of preferences that puts a negative weight on the monetary payoff of the other player (such as competitive preferences or envy) removes the payoff dominant equilibrium and renders the sabotage equilibrium (AA) unique. On the other hand, preferences that put some non-negligible weight on either the payoff of the other player or on total surplus remove the sabotage equilibrium and render the efficient equilibrium (BB) strict and unique. [5] Self-centered inequality aversion in the formulation of Fehr and Schmidt (1999) increases the payoff differences between equilibrium and out-of-equilibrium play, which according to Battalio et al. (2001) favours the risk-dominant sabotage equilibrium.[6]

[5]For the sabotage equilibrium to disappear, the player needs to be willing to sacrifice one unit of her own payoff for either the other player gaining 39 units or for the total surplus increasing by 38 units.

[6]This is also the case in psychological games with reciprocity and fairness motives such as Rabin (1993).

2.2 Treatments and Hypothesis

Recall that the purpose of our design was to investigate whether the existence of bonuses leads to more efficiency damaging sabotage behaviour even if the bonuses do not provide material incentives for sabotage. In our baseline game without a bonus, the sabotage action does not change the payoff of the saboteur conditional on the other player not sabotaging. Hence, there is no incentive to sabotage unilaterally. In order to isolate the effect of the introduction of a bonus without a financial incentive, we need to keep the effective payoffs of the game the same but introduce a bonus. This can be achieved by introducing a bonus of $b < s$, for the player who ends up with the higher payoff, and at the same time introducing a cost of unilateral sabotage $c = b$ that offsets the bonus.

In the control treatment, players play the base game with a slight teamwork frame. Players are referred to as "teammates".[7] In our two experimental treatments participants play the payoff equivalent game with a bonus of $b = 30$ for the player who plays A if the other player plays B. In that case, the bonus is offset by an equivalent sabotage cost of $c = 30$. The two experimental treatments differ with respect to the framing. In the bonus frame (see the screenshot in Figure 1), we maintain the team frame and frame the bonus neutrally by stating that "if one person chooses A and the other person chooses B, then the person who chose A gets an extra 30 points." The difference between the control treatment and the bonus treatment is the introduction of a pay-off-invariant bonus.

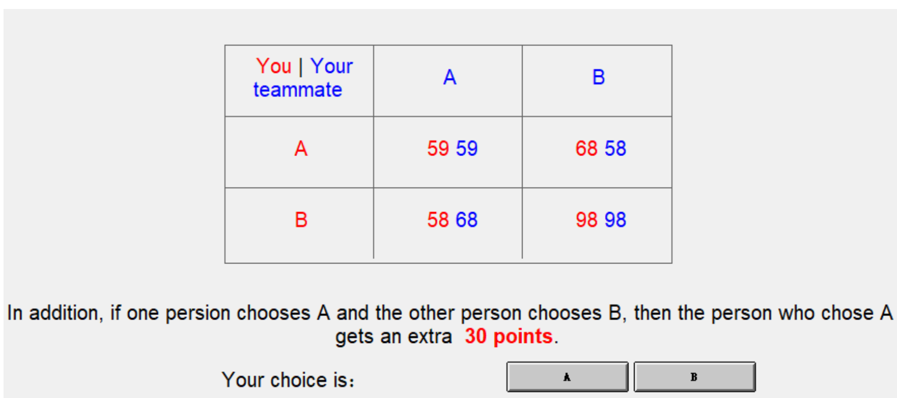


Figure 1: Screenshot in the bonus treatment

[7] Note that all wording presented here is translated from Mandarin.

Our second experimental treatment uses the version of the game with a bonus from the bonus treatment and adds a competitive frame to it. Players are now referred to as "rivals", and the condition for being paid the bonus is framed in terms of a payoff comparison: "the person who gets more points than the other will win an extra 30 points as a bonus." See Figure 2 for a screenshot.

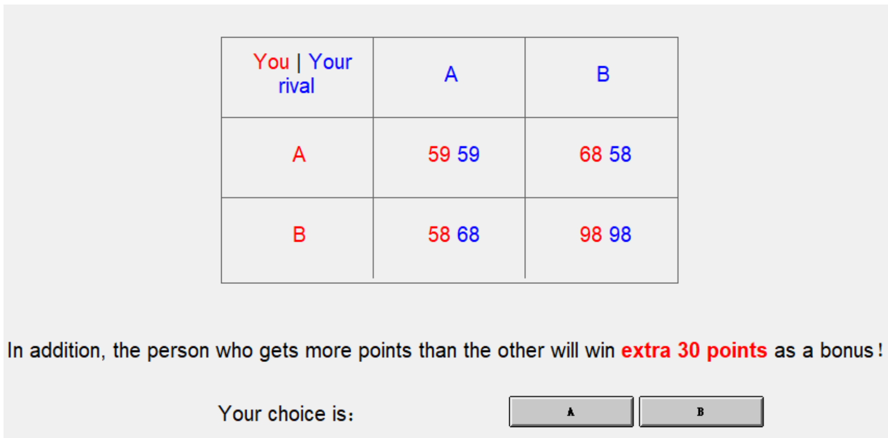


Figure 2: Screenshot in the competitive bonus treatment

2.3 Implementation

The experiment consists of three treatments, i.e. *control*, *bonus* and *competitive bonus*. We repeated the game for 20 periods and implemented a repeated single-shot design where participants were randomly re-matched within a group of four people.[8] At the end of each round, we provided feedback about the outcome of the game in line with the framing condition. In both the control and bonus treatment, participants are informed about their own and teammate's choices and points earned. In the competitive bonus treatment, participants are informed about their and their rival's choices and points earned. In addition, they are also informed about the outcome with respect to the bonus, i.e. if anybody got more points than the rival and therefore was paid a bonus.

Each session consisted of an even number of subjects between 24 and 40 participants. We conducted experiments in December 2021 and May 2022 at the Research Center for

[8]Participants were informed that they would be randomly re-matched by the computer before the beginning of each period game.

Behavioural Science of Wuhan University in China. Participants were recruited through ancademy.org (an online application that uses Wechat, the dominant chatting app in China, for communication). A total of 394 university students from a variety of disciplines participated. We use the same neutral framing in identical written instructions to explain the basic game structure (simultaneous, symmetric 2x2 game) in all treatments. The experimenter first read the instructions, and then the participants had to successfully answer some control questions to ensure they understood the normal form of the game before the start of the experiment. Treatments were programmed in Z-Tree (Fischbacher, 2007). Treatment differences only occurred in the computerized instructions that followed. There, participants were informed of their role as "teammates" (control and bonus treatment) or "rivals" (competitive bonus treatment) in the upcoming game and participants saw the payoff values and the description of the bonus if applicable), as detailed in the figures above. Participants earned points that were converted at the end of the session at an exchange rate of 1 RMB for 50 points. In addition, participants received a show-up fee of 10 RMB. We paid on average roughly 35 RMB for a one-hour session into the participants' ancademy account. The payment occurred within 48 hours of the participation.

The central question of this paper is whether the existence of bonuses can lead to more efficiency-damaging sabotage behaviour even if they do not provide explicit financial incentives. Note that both players choosing not to sabotage is the socially efficient and only Pareto optimal outcome. One of the two players sabotaging leads to the next highest surplus, while both sabotaging yields the lowest surplus. Given that the underlying financial incentives are identical in all three treatments, a natural null hypothesis is that play, and hence the level of efficiency does not differ across treatments. Alternatively, the framing with a bonus might induce more people to sabotage they value a payoff of 68 plus a bonus of 30 more than a payoff of 98 without a bonus. Moreover, framing the bonus as the outcome of having a higher payoff than a rival might, in addition, activate competitive preferences. Under this alternative hypothesis, we would expect the average payoff to differ between treatments with payoffs being largest in the control than in the bonus and larger in the bonus than in the competitive bonus treatment.

3 Results

There are 30, 33, and 34 independent matching groups in the control, bonus, and competitive bonus treatments, respectively.[9] In what follows, we present our main results and then dig into the mechanism behind it.

3.1 Efficiency

We first investigate whether the existence of a bonus leads to a reduction in efficiency. The joint payoff reaches the highest level if both players do not sabotage, an intermediate level if one player sabotages and the other does not, and the lowest level if both players sabotage. We measure efficiency by the average payoff per person. An independent observation is the average profit taken over all players within a matching group and over all periods. Figure 3 describes the mean value with a 95% confidence interval.

The average efficiency is 92.24 in the control treatment, 88.71 in the bonus treatment, and 85.00 in the competitive bonus treatment. The efficiency drop from the control treatment to the bonus treatment and from the bonus treatment to the competitive bonus treatment is weakly significant (one-sided t-test, $p < 0.06$ and $p < 0.08$, respectively), and the efficiency gap between the control and competitive bonus treatments is strongly significant (one-sided t-test, $p < 0.001$).[10] The finding supports our initial conjecture that the existence of bonuses destroys efficiency, even though the bonus provides no financial sabotage incentives. This effect is particularly strong if the bonus comes with a competitive framing.

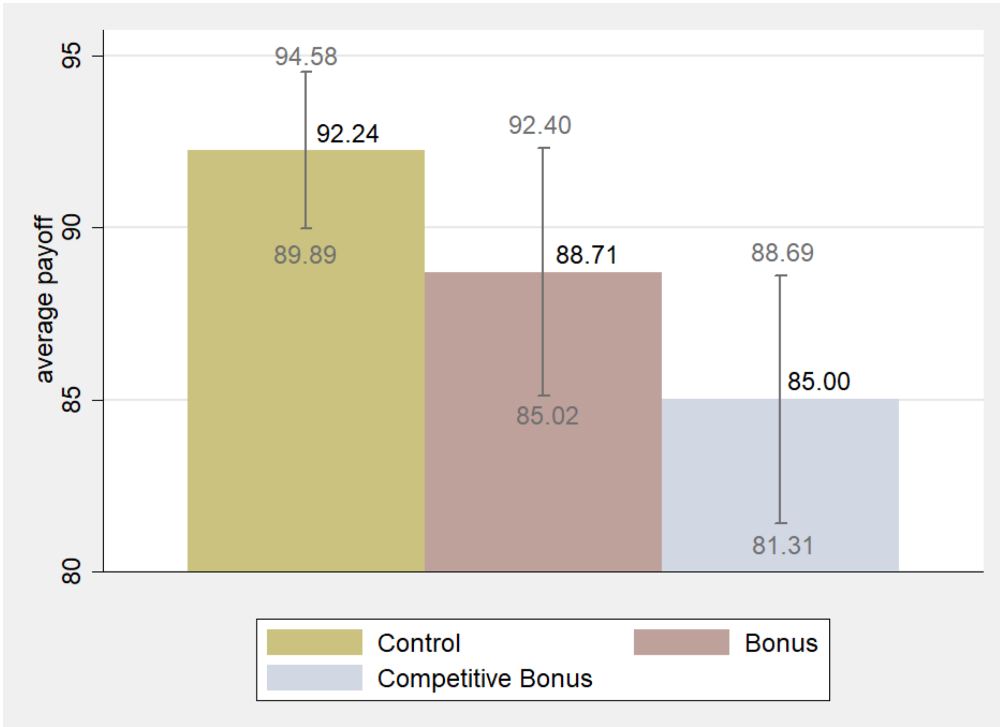
Result 1 *Bonuses framed as the outcome of a competition with rivals harm efficiency.*

Next, we explore the mechanism behind this treatment effect. For this, it is instructive to compare the relative frequencies of the three different outcomes between treatments. Table 3 reports the relative frequencies by treatments separately for the first, last and intermediate periods. This separation takes into account that we observe a) relatively similar distributions of outcomes across treatments in period 1, b) very distinct dynamics

[9] In three sessions we had two extra participants each who were paired for the whole experiment. These six participants were removed from our analysis because random matching was no longer applied.

[10] We use a t-test since we have more than 30 observations per treatment, which as a rule of thumb guarantees approximate normal distributions of sample means. Permutation and rank sum tests yield very similar results.

Figure 3: Average profit across treatments



between periods 1 and 2, c) stable but different fractions of the efficient outcome in periods 2 to 19, and d) an endgame effect in period 20.

Table 3: Relative frequencies in percent of outcomes

Treatments	First Period			Periods 2-19			Last Period		
	AA	**BB**	AB,BA	AA	**BB**	AB,BA	AA	**BB**	AB,BA
Control	0.00	**68.33**	31.67	3.52	**75.93**	20.56	11.67	**60.00**	28.33
Bonus	1.52	**74.24**	24.24	11.87	**64.81**	23.32	24.24	**57.58**	18.18
Comp Bonus	4.41	**75.00**	20.59	18.30	**51.31**	30.39	22.06	**39.71**	38.24

The driver of the aggregate efficiency differences is the stable outcome differences in periods 2 to 19. The relative frequencies of the two equilibrium outcomes are markedly different. In the control treatment, virtually all equilibrium play concentrates on the efficient equilibrium without sabotage, i.e.BB. The efficient equilibrium was played 21.6 times more often than the inefficient mutual sabotage equilibrium. Adding a bonus without changing the framing reduces the relative occurrence of BB compared to AA to a ratio of 5.5 to 1. Adding the competitive framing further reduces the ratio to 2.8 to 1.

Result 2 *Both the introduction of a bonus and a competitive framing shift play from the efficient to the inefficient equilibrium.*

The relative frequency of intermediate, non-equilibrium outcomes where one participant sabotages, while the other does not, i.e. (AB, BA), is somewhat higher in the competitive bonus treatment than in the other two treatments. However, splitting the sample into 5-period intervals and comparing the relative frequencies (see Table 4) does not reveal any obvious differences between treatments. This is further evidence that the main driver of efficiency damages through bonuses is a shift from the efficient to the inefficient equilibrium.

Table 4: Mismatch fraction across treatments

Treatments	periods 1-5	periods 6-10	periods 11-15	periods 16-20
Control	23.33	21.33	19.67	21.67
Bonus	26.97	23.33	23.64	18.48
Competitive Bonus	28.35	33.53	27.94	25.59

3.2 Dynamics of play

The weak dominance of the sabotage strategy A for purely selfish preferences suggests that learning should lead to a very gradual shift of play towards the inefficient equilibrium. Different kinds of social preferences shift dominance to the non-sabotage strategy B (e.g., altruism or a taste for efficiency), or further strengthen the attraction of the sabotage strategy A (e.g., envy or competitive preferences).

To see this, consider social preferences of the form $U_i(\Pi_i, \Pi_{-i}) = \Pi_i + \alpha\Pi_{-i}$, where Π_i is the own payoff and Π_i is the material payoff of the other player. The parameter α measures how much player i cares about the payoff of the other player. For standard selfish preferences, we have $\alpha = 0$. A positive α reflects a warm glow toward the other player. A negative value represents competitiveness. The payoff differences between playing A and

B conditional on the other players choice are given by

$$U_i(A, A) - U_i(B, A) = (59 + \alpha 59) - (58 + \alpha 98) = 1 - 39\alpha \quad (1)$$

$$U_i(A, B) - U_i(B, B) = (98 + \alpha 58) - (98 + \alpha 98) = -40\alpha \quad (2)$$

For competitive preferences ($\alpha < 0$) both differences are positive, and sabotaging (playing A) becomes a dominant strategy. Team-oriented preferences ($\alpha > 0$) make the non-sabotage equilibrium strict, since not sabotaging (playing B) is now strictly preferred to sabotaging if the other player does not sabotage as $U_i(A, B) - U_i(B, B) < 0$. Moreover, for a very modest warm glow of ($\alpha > 1/39$) not sabotage becomes the dominant strategy, and the equilibrium (B, B) becomes unique.

If our different treatments activate different social preferences, then we would expect the dynamics of play to evolve differently between treatments. Figure 4 shows the fraction of matching groups per treatment in which all four players do not sabotage or where all players do sabotage. This is instructive, as we expect that over time through interaction in the majority of matching groups, one of the two equilibria would prevail. In fact, in the latter half of the experiment, more than half of the matching groups in all treatments exhibit uniform play.

Treatment differences arise from different but relatively stable fractions of non-sabotage matching groups and from the degree to which overtime matching groups where everybody sabotages develop. The control treatment without bonus has the highest number of matching groups without any sabotage and virtually no matching groups drift into full sabotage.[11] Introducing a bonus (without the competitive framing) slightly reduces the number of non-sabotage groups, and late in the experiment a significant number of matching groups where everybody plays A .[12] In the treatment with a bonus and the additional competitive framing, there are the least fully cooperative groups, and the fraction of matching groups where all four participants choose the sabotage strategy rises steadily.[13]

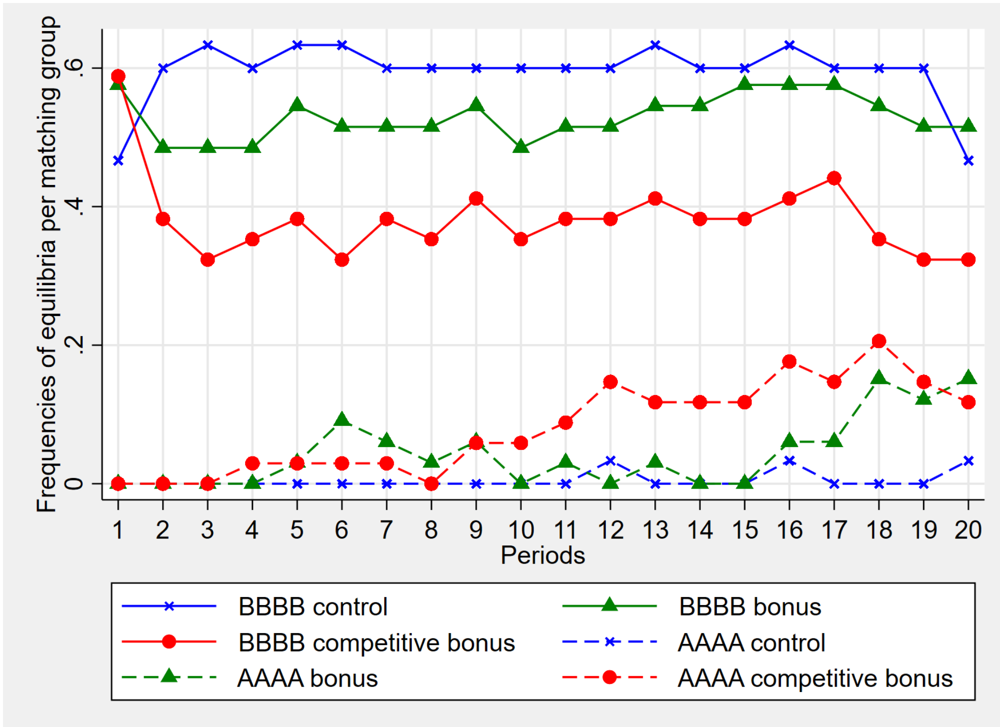
Result 3 *After the initial period, the fractions of matching groups where nobody sabotages*

[11] According to a t-test with clustering on a matching group the fraction of $BBBB$ is significantly higher in the Control treatment than in the Competitive Bonus Treatment ($p < 0.016$, one-sided).

[12] The difference in the fraction of $BBBB$ is not significantly different though ($p > 0,25$, one-sided t-test adjusted for clustering. However, the fraction of play of $AAAA$ is significantly greater in the Bonus treatment than in Control ($p < 0.016$, one-sided)

[13] $BBBB$ is weakly significantly less frequent than in the Bonus treatment ($p < 0.075$, one-sided), and $AAAA$ is significantly more frequent than in the Control ($p < 0.016$, one-sided).

Figure 4: Dynamics of equilibria within matching groups across treatments



are stable and lowest in the competitive bonus treatment.

Result 4 *In both bonus treatments over time, the fraction of uniformly sabotaging matching groups rises, while in the control treatment fully sabotaging matching groups are extremely rare (with a relative frequency of* 0.005*).*

From Figure 4 we can see that the main difference in the stable fractions of fully cooperating matching groups is due to the behavioural adjustments between periods 1 and 2. It is instructive to investigate how the likelihood that a subject plays B (that is, not sabotage) in period 2 conditional on her experience in period 1 differs across treatments. Table 5 shows the percentages of the play of B conditional on the observed history. The first entry is the player's own strategy in period 1, while the second entry is the action played by the other player.

Table 5: Percentage of $S_i = B$ in $t = 2$ conditional on the outcome in $t = 1$.

$(S_i S_{-i})_{t=1}$	Control	Bonus	Com Bonus
$(AA)_{t=1}$	–	50.00	66.67
$(BB)_{t=1}$	96.34	92.86	90.20
$(AB)_{t=1}$	57.89	50.00	**14.29**
$(BA)_{t=1}$	**94.74**	56.25	42.86

The predominant difference between treatments is the reaction of players to disequilibrium play. In disequilibrium play, one player sabotages while the other does not. The sabotaging player receives a payoff of 98 and the non-sabotaging player earns only 58. Players who think about what would have happened if I had chosen the other strategy realize that the main difference would have been the payoff of the other person. If the saboteur had not sabotaged her payoff would have stayed the same, but the other player would have received 30 units more. The person who did not sabotage would have increased her own payoff by 1 and reduced the payoff of the other person by 29 if she had chosen to sabotage. The attractiveness of either switch depends on the social preferences of a player. Preferences putting a positive weight on the payoff of the other player (such as preferences for social efficiency or altruism) make it attractive to switch away from sabotage or not to switch to sabotage depending on the previous play. Competitive preferences make sticking or switching to sabotage more attractive.

In the control treatment, almost all subjects who experienced sabotage without sabotaging themselves chose to refrain from sabotaging in the next period, while in the bonus treatments, only about half of the subjects with this experience stuck to choosing the friendly strategy B. Unilateral saboteurs, predominantly (more than 85 percent) stuck to sabotage in the competitive bonus treatment, while in the other two treatments, at least half of them switched to the kind strategy B.

We conjecture that the initial experience of off-equilibrium play, for some participants, made the framing (team, bonus, or bonus if you have a higher payoff) more salient and activated the corresponding social preferences. This then led to a very different change between the fraction of matching groups that were able to fully prevent sabotage between periods 1 and 2. This is plausible if we take into account the different feedback subjects get. After being the sole saboteur you are reminded in the competitive bonus treatment that you got a bonus because your payoff was higher than the payoff of the rival. This reinforces trying to get a higher payoff than the competitor. In both the control and the bonus treatment. There is more emphasis that you damaged the payoff of your teammate. Similarly, if you have been the target of unilateral sabotage, then the combination of having a low payoff and being told that the other person got a bonus (in both bonus treatments) makes it clear to you that you are the sucker. Not sticking with B ensures that you do not become the sucker in the next period again. In contrast, in the control treatment, there is no mention of any bonuses. Hence, you do not feel like the sucker, but the generous person responsible for a high payoff of your teammate. There is no particular urge to change this.

Finally, we look at the dynamics of the outcomes for matching groups in periods 2 to 20, where the fraction of sabotage-free matching groups is stable in all three treatments. Table 6 reports the Markov transition probabilities (in percent). Each matching group can be in one of the six states $S_t \in \{(AA|AA), (AA|AB), (AB|AB), (AA|BB), (AB|BB), (BB|BB)\}$ in each period. The states in the rows are the origin states while the columns contain the destinations. The table entries give the estimated time-invariant probability that a matching group transitions to S_{t+1} conditional on having started in state S_t. We also calculated the limiting distribution of states which we denote by π. This is the distribution of states to which matching groups will converge given the transition probabilities.[14]

We highlight the most interesting differences between the treatments by printing the entries in boldface. The two most important differences in the dynamics between treatments are differences in the persistence of states in which all members of the matching group sabotage $(AA|AA)$ or all refrain from sabotaging $(BB|BB)$. The probability of staying sabotage-free conditional on having been sabotage-free the period before is lower in the competitive bonus treatment (89 percent vs 96 and 94 percent). More dramatically, in the control treatment, no group that ended up with only sabotage choices remained that way in the next period. In the bonus treatment, the probability was more than half (54 percent). In the competitive bonus treatment, the probability of a group with only sabotage being only sabotage again in the next period was greater than 2/3 (71 percent).

[14] Note that in our estimated Markov matrices, matching groups have a path to any possible state regardless of the initial state, which implies that the limiting distribution is unique and independent of the initial conditions.

Table 6: Transition probabilities of matching groups between outcomes from period 2 to 20.

		$\mathbf{S_{t+1}}$						
	Control	$(AA\|AA)$	$(AA\|AB)$	$(AB\|AB)$	$(AA\|BB)$	$(AB\|BB)$	$(BB\|BB)$	π
	$(AA\|AA)$	**0.00**	50.00	0.00	0.00	50.00	0.00	**1.00**
	$(AA\|AB)$	8.33	33.33	33.33	25.00	0.00	0.00	4.50
	$(AB\|AB)$	5.71	14.29	22.86	22.86	31.43	2.86	10.23
	$(AA\|BB)$	0.00	9.09	31.82	27.27	27.27	4.55	7.07
	$(AB\|BB)$	0.00	1.56	14.84	6.25	68.75	8.59	26.83
	$(BB\|BB)$	0.00	0.00	0.32	0.00	5.47	**94.21**	**50.41**
	Bonus	$(AA\|AA)$	$(AA\|AB)$	$(AB\|AB)$	$(AA\|BB)$	$(AB\|BB)$	$(BB\|BB)$	π
	$(AA\|AA)$	**54.17**	41.67	4.17	0.00	0.00	0.00	**8.60**
	$(AA\|AB)$	22.95	59.02	9.84	3.28	4.92	0.00	15.66
$\mathbf{S_t}$	$(AB\|AB)$	4.17	27.08	27.08	22.92	18.75	0.00	8.27
	$(AA\|BB)$	0.00	3.33	43.33	16.67	30.00	6.67	4.81
	$(AB\|BB)$	0.00	2.97	12.87	10.89	63.37	9.90	14.69
	$(BB\|BB)$	0.00	0.00	0.34	0.00	3.37	**96.30**	**41.20**
	Comp Bonus	$(AA\|AA)$	$(AA\|AB)$	$(AB\|AB)$	$(AA\|BB)$	$(AB\|BB)$	$(BB\|BB)$	π
	$(AA\|AA)$	**70.59**	21.57	1.96	5.88	0.00	0.00	**13.27**
	$(AA\|AB)$	17.65	47.06	21.18	5.88	8.24	0.00	17.91
	$(AB\|AB)$	5.80	40.58	23.19	11.59	15.94	2.90	12.79
	$(AA\|BB)$	0.00	14.71	47.06	11.76	23.53	2.94	5.74
	$(AB\|BB)$	0.00	3.25	15.45	8.94	55.28	17.07	17.91
	$(BB\|BB)$	0.00	0.00	0.93	0.46	9.72	**88.89**	**32.38**

Under the assumption that the dynamics remain stable, the state distributions of the matching groups converge to the limiting distribution (π in the table). This distribution can be seen as the long-run outcome of the dynamics between periods 2 and 20. The results show that given the observed dynamics, we expect treatment differences to persist. The stable fraction of fully sabotaging groups is lowest in the control treatment (1 percent), higher in the bonus treatment (8.6 percent) and highest in the competitive bonus treatment (13.3 percent). The stable fraction of sabotage-free matching groups ranks the opposite way. We have the highest fraction in the control treatment (50.4 percent), a lower fraction in the bonus treatment (41.2 percent) and the lowest fraction in the competitive bonus treatment (32.4 percent).

These persistent differences in matching-group outcomes imply long-term efficiency differences. We calculated the expected surplus per person implied by the outcome distributions. With an expected per-person surplus of 90.18 currency units, the control treatment is the most efficient, with 85.97 units the bonus treatment is somewhat less efficient, and the competitive bonus treatment is least efficient with 82.07 units. A comparison with the average surplus over the whole experiment shows that the long-run efficiency levels are lower by 2 to 3 percentage points and that the magnitude of the treatment differences is persistent and slightly larger.

Result 5 *Conditional on the observed matching-group dynamics, the efficiency damaging impact of bonuses with and without competitive framing is persistent.*

4 Conclusion

In many environments with competitive bonuses and prizes, participants have material incentives to sabotage others. In this paper, we presented experiments that explore the impact of bonuses or prizes on sabotage activity even if they do not give rise to material sabotage incentives. We find that efficiency-damaging sabotage activity becomes more prominent when we introduce a bonus into a sabotage game, without changing the material payoffs and the framing as a team game. The effect becomes stronger when the bonus is accompanied by a competitive framing that emphasizes the social comparison and feedback of competitiveness. We conclude that using bonuses and prizes can lead to more sabotage behaviour even if their introduction does not provide material incentives to do so. Our base game and setup are such that we can be confident that the mechanism leading to the increased sabotage activity is due to bonuses and the competitive framing shifting social preferences toward competitiveness. Hence, we expect that our results that were obtained from a very specific environment generalize to other settings where bonuses and prizes are present since more competitive preferences are conducive to more sabotage behaviour in any environment.

References

Alfitian, J., Sliwka, D., and Vogelsang, T. (2023). When bonuses backfire: Evidence from the workplace. *Management Science*.

Battalio, R., Samuelson, L., and Van Huyck, J. (2001). Optimization incentives and coordination failure in laboratory stag hunt games. *Econometrica*, 69(3):749–764.

Burks, S., Carpenter, J., and Goette, L. (2009). Performance pay and worker cooperation: Evidence from an artefactual field experiment. *Journal of Economic Behavior & Organization*, 70(3):458–469. Field Experiments in Economics.

Buser, T. and Dreber, A. (2016). The flipside of comparative payment schemes. *Management Science*, 62(9):2626–2638.

Carpenter, J., Matthews, P. H., and Schirm, J. (2010). Tournaments and office politics: Evidence from a real effort experiment. *The American Economic Review*, 100(1):504–517.

Charness, G., Masclet, D., and Villeval, M. C. (2014). The dark side of competition for status. *Management Science*, 60(1):38–55.

Chen, K. (2003). Sabotage in Promotion Tournaments. *The Journal of Law, Economics, and Organization*, 19(1):119–140.

Dal Bó, P., Fréchette, G. R., and Kim, J. (2021). The determinants of efficient behavior in coordination games. *Games and Economic Behavior*, 130:352–368.

Deutscher, C., Frick, B., Gürtler, O., and Prinz, J. (2013). Sabotage in tournaments with heterogeneous contestants: Empirical evidence from the soccer pitch. *The Scandinavian Journal of Economics*, 115(4):1138–1157.

Drago, R. and Garvey, G. T. (1998). Incentives for helping on the job: Theory and evidence. *Journal of Labor Economics*, 16(1):1–25.

Dubois, D., Willinger, M., and Van Nguyen, P. (2012). Optimization incentive and relative riskiness in experimental stag-hunt games. *International Journal of Game Theory*, 41(2):369–380.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation*. *The Quarterly Journal of Economics*, 114(3):817–868.

Fischbacher, U. (2007). Z-tree - Zurich toolbox for readymade economic experiments. *Experimental Economics*, 10(2):171–178.

Flory, J. A., Leibbrandt, A., and List, J. A. (2016). The Effects of Wage Contracts on Workplace Misbehaviors: Evidence from a Call Center Natural Field Experiment. NBER Working Papers 22342, National Bureau of Economic Research, Inc.

Gangadharan, L., Grossman, P. J., and Vecci, J. (2020). Antisocial behavior in the workplace. *Handbook of Labor, Human Resources and Population Economics*, pages 1–26.

Gill, D., Prowse, V., and Vlassopoulos, M. (2013). Cheating in the workplace: An experimental study of the impact of bonuses and productivity. *Journal of Economic Behavior & Organization*, 96:120–134.

Gläser, D., van Gils, S., and Van Quaquebeke, N. (2017). Pay-for-performance and interpersonal deviance: Competitiveness as the match that lights the fire. *Journal of Personnel Psychology*, 16(2):77–90.

Gläser, D., van Gils, S., and Van Quaquebeke, N. (2022). With or against others? pay-for-performance activates aggressive aspects of competitiveness. *European Journal of Work and Organizational Psychology*, 31(5):698–712.

Gonzalez, V., Dalton, P. S., and Noussair, C. (2020). The dark side of monetary bonuses: Theory and experimental evidence.

Grund, C. and Sliwka, D. (2005). Envy and compassion in tournaments. *Journal of Economics & Management Strategy*, 14(1):187–207.

Harbring, C. and Irlenbusch, B. (2005). Incentives in tournaments with endogenous prize selection. *Journal of Institutional and Theoretical Economics (JITE)*, 161:636–636.

Harbring, C. and Irlenbusch, B. (2008). How many winners are good to have?: On tournaments with sabotage. *Journal of Economic Behavior & Organization*, 65(3):682–702.

Harbring, C. and Irlenbusch, B. (2011). Sabotage in tournaments: Evidence from a laboratory experiment. *Management Science*, 57.

Harsanyi, J. C., Selten, R., et al. (1988). A general theory of equilibrium selection in games. *MIT Press Books*, 1.

Kandori, M., Mailath, G. J., and Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56.

Lazear, E. P. (1989). Pay equality and industrial politics. *Journal of political economy*, 97(3):561–580.

Lazear, E. P. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of political Economy*, 89(5):841–864.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.