Bad News: An Experimental Study On The Informational Effects Of Rewards*

Andrei Bremzen†, Elena Khokhlova‡, Anton Suvorov† and Jeroen van de Ven§

April 6, 2012

Abstract

Both psychologists and economists have argued that rewards often have hidden costs. One possible reason is that the principal may have incentives to offer higher rewards when she knows the task to be difficult. Our experiment tests if high rewards embody such bad news and if this is perceived by their recipients. Our design allows us to decompose the overall effect of rewards on effort into a direct incentive and an informational effect. The results show that most participants correctly interpret high rewards as bad news. In accordance with theory, the negative informational effect co-exists with the direct positive effect.

Keywords: reward, bonus, informational content, motivation, crowding-out, laboratory experiment.

JEL codes: D82, D83, J33

*We are grateful to Akhmed Akhmedov, Alex Koch, Grigory Kosenok, Ruben Enikolopov, and participants of the EMIR workshop in Lyon (2011) for helpful comments and suggestions.

†CEFIR and New Economic School, Nakhimovsky Prospekt, 47, Suite 1721, 117418, Moscow, Russia.

‡McKinsey&Company, 5 Lesnaya St., Building "C", 125047, Moscow, Russia.

§Corresponding author. University of Amsterdam, ACLE, Roetersstraat 11, 1018 WB Amsterdam, the Netherlands. Email: j.vandeven@uva.nl

1 Introduction

Rewards are used in many types of relationships. While there is much evidence that rewards can be an effective way of motivating people, there is also a vast collection of experiments showing that rewards can have unintended consequences. Often, these negative effects of rewards are hidden at first, and do not manifest themselves until later in the relationship. For instance, the promise of a gift for obtaining high grades at school may well keep a child studying hard, whilst at the same time undermining any genuine interest in learning and thereby having profound negative consequences later on. Similarly, promising a gratification to employees for successfully completing a project may well temporarily increase their efforts, only to result in a reduced interest in their job afterwards. A good understanding of why and when such negative effects are most likely to occur is important for the optimal design of contracts and other incentive schemes.

We conducted an experiment to bring these hidden costs to the surface. We study an environment in which the principal has incentives to promise a higher bonus when she knows that the task is difficult. We find that agents understand this, and interpret the bonus as bad news. This negative *information effect* induces costs that are usually hidden because in the short term they are outweighed by the direct positive *incentive effect*. Our experimental design allows to decompose the overall impact on motivation into these two different effects, a feature that distinguishes our experiment from the existing literature.

In our experiment, two players are anonymously matched to each other, one in the role of the principal ("she"), the other in the role of the agent ("he"). The design is based on a simplified version of the model by Bénabou and Tirole (2003) which gives a game-theoretic explanation for the "hidden costs" effect based on information asymmetries. A key element is that the agent is uncertain about the task difficulty (i.e., cost of effort), while the principal knows whether the task is easy or difficult. In the first stage, the principal decides upon an up-front fixed wage and a bonus that is contingent on good performance. In the second stage, after observing the bonus and the wage, the agent chooses whether or not to exert effort. Good performance

requires exerting effort, and results in a higher joint profit of the players irrespective of the task difficulty. Parameters are such that without a bonus, the agent would gain from exerting effort on the easy task, which is sufficiently self-rewarding, but not on the difficult task. In equilibrium the principal offers a bonus only when she observes high costs. Thus, a high reward increases effort but brings bad news for the agent, resulting in potential hidden costs.

The key feature we introduced in the experimental design is an additional project for the agent. Besides the *joint project* with the principal, the agent also chooses an effort level for his *own project*. The only difference between the projects is that the bonus and the wage specified by the principal do not apply to the agent's own project. This takes away the incentive effect of the bonus, but not the information effect, and therefore allows us to isolate the informational content as perceived by the agent.

The results provide clear support for the main predictions of the model. First, we find that the bonus offered by the principal is strongly related to the difficulty of the project in the informed condition: when costs are high, the principal is 50 percentage points more likely to give a bonus. Thus, the bonus is very informative about the cost level, and the principal understands the need to offer a high reward when costs are high. Secondly, a high bonus is very effective in stimulating effort in the joint project through the direct incentive effect (the monetary benefits of the reward). Finally, we also find evidence of the informational effect of rewards (the hidden costs): rewards are correctly perceived by the agents as conveying bad news, decreasing their motivation to invest in their own project. This effect becomes especially strong in later rounds. In the last 10 rounds, the likelihood of the agents' exerting high effort on their own project is around 34 percentage points lower after receiving a bonus.

To investigate the agents' reaction to bonuses that have no informational content we also introduced a control treatment, in which the principal had no private information about the task. As predicted, we find that in the control treatment a bonus is still very effective in stimulating effort in the joint project, but the negative effect on effort in the own project is mostly absent. A possible concern might furthermore

be that a high bonus signals the principal's altruistic attitude rather than task difficulty. Therefore, as a further control, we also elicited some components of social preferences of participants using various modifications of a trust game. We do not find any support that the above results are driven by fairness considerations.

This paper is related to a vast literature that explores "crowding out" of intrinsic motivation by rewards or other types of extrinsic incentives.[1] Experiments in social psychology, starting from Deci (1971), Kruglanski et al. (1971) and Lepper et al. (1973), have shown that a promise of a performance-contingent reward for an interesting task may undermine a participant's attitude to the task and make his or her future engagement in similar activities less likely in the absence of rewards. This long-term negative effect (the hidden costs) may coexist with the immediate positive effect of rewards that act as short-term reinforcers. Two types of arguments have been put forward for explaining such effects. The first emphasizes the controlling aspect of rewards. Rewards undermine participants' self-determination to engage in the task and do the task well (see Deci and Ryan (1985)). The other underscores the informational aspects of rewards: agents perceive high rewards as embodying bad news about task difficulty and their ability to complete the task successfully. This interpretation of rewards comes from the "overjustification effect", according to which people start to attribute their engagement in any activity to the external rewards, displacing part of their intrinsic interest. In psychology these ideas can be accommodated by theories based on cognitive dissonance (Festinger (1957)) or, alternatively, on self-perception theory (Bem (1967)). Bénabou and Tirole (2003) explore this idea in a game-theoretic framework and show that these hidden costs can indeed occur as an equilibrium phenomenon.

Of course, agents can only make proper inferences from rewards if they are aware of the principals' objectives. In Deci (1971) and related experiments, however, rewards have been administered by the experimenter, whose objectives were not clear to participants.[2] To the best of our knowledge ours is the first experiment in which

[1] See Frey (1997), Frey and Jegen (2001) and Fehr (2002) for a discussion of many earlier contributions.

[2] See a meta-analysis in Deci et al. (1999) or a book Deci and Ryan (1985) for extensive accounts of this literature; see also Lepper et al. (1999) and Eisenberger et al. (1999) for a different perspective.

rewards are determined by active *participants* with well-defined objectives that are common knowledge to all participants, and the information asymmetry about the task is directly introduced into the experiment in a controlled manner.

Another important strand of literature demonstrates crowding-out effects in experimental labor markets, often using variations of the gift-exchange game by Fehr et al. (1993) and Fehr et al. (1997). In contrast to our work, though, all these studies are focused on how extrinsic incentives interact with various aspects of broadly defined social preferences. In particular, Fehr and Gächter (2001) show that the use of both performance-contingent rewards and sanctions reduces effort provision and aggregate payoffs (see also Fehr and List (2004)). Fehr and Schmidt (2007) show that adding a stick (a fine) to a carrot (a bonus) in an incentive contract may have a detrimental effect on the agents' performance. Relatedly, in a modified trust game where the investor has an option to impose sanctions on the trustee for insufficient cooperation, Fehr and Rockenbach (2003) show that using the option to fine the trustee backfires compared to a pure trust game where this option is unavailable. In contrast, withdrawing from applying this option when it is available has a positive impact both on the aggregate and on the principal's own average payoff. An explanation put forward in these experiments is that the principal's reliance on extrinsic incentives or control signals her lack of trust in the agent, who then reciprocates by indeed behaving in a distrustful manner.[3]

In a field experiment Gneezy and Rustichini (2000a) find that the introduction of a fine on parents that arrive late to collect their children at a day-care increases the occurrences of late-coming parents, rather than deter parents from doing so. They interpret this effect in terms of learning by the parents about the mildness of the day-care owners. Ariely et al. (2009) find a detrimental effect on performance when rewards become very high, consistent with the idea that people experience increased arousal and choke under the pressure. In contrast, Gneezy and Rustichini (2000b) show that very small performance-contingent rewards impair their performance com-

[3]Sliwka (2007) investigates in a theoretical model how information about social norms of behavior can be transmitted from more informed principals to less informed agents via the choice of incentive schemes.

pared to no-reward condition, possibly because they insult the agents.

Several experimental studies show that other types of interventions can have a detrimental impact on performance. Falk and Kosfeld (2006) showed that the principal's choice to control the agent (i.e., enforce a minimum effort) reduces the agents' performance because most agents perceive control as a signal of distrust and low expectations by the principal. Galbiati et al. (2009) examine the effects of sanctions in a coordination game. Cooperative subjects perceive endogenous sanctions by a third party as a negative signal about the contributions of others, which takes away the sanction effect. Relatedly, Charness et al. (2010) showed that delegating the wage choice to agents increases effort. Dickinson and Villeval (2004) study the relation between the degree of monitoring and effort, finding some support for crowding out.

The rest of the paper is organized as follows. In the next section, we present the model. In section 3 we describe the experimental setup and hypotheses. The results are described in section 4, and the final section concludes.

2 The Model

2.1 Informed Principal

The main treatment of our experiment is based on a simplified version of the model by Bénabou and Tirole (2003). There are two risk-neutral players, a principal (she) and an agent (he). The agent works on a task that is potentially self-rewarding. He chooses a binary effort level, $e \in \{0, 1\}$. The low level of effort, $e = 0$, implies no cost, and leads to payoffs A_0 and P_0 for the agent and the principal respectively. The high level of effort, $e = 1$, costs $c > 0$ to the agent. It results in a higher output, and yields an additional payoff of $\Delta A > 0$ for the agent and $\Delta P > 0$ for the principal. To stimulate the agent, the principal may promise a bonus b to be paid if the agent chooses the high effort level. Thus, her payoff is:

$$U^P = P_0 + e(\Delta P - b),$$

where $b \in \{0, \overline{b}\}$. The agent's payoff is:

$$U^A = A_0 + e(\Delta A + b - c).$$

There is uncertainty about the cost of effort: it is common knowledge that c is equally likely to be high, c_H, or low, $c_L < c_H$. This can be interpreted as uncertainty about the difficulty of the task. The principal is perfectly informed about the difficulty of the task. The agent only has a rough idea about the level of costs: he receives a private signal, s, about the cost of effort which assumes two possible values, $s \in \{s_L, s_H\}$. With probability $r > 0.5$ the signal is correct, i.e., signal s_i arrives when costs are c_i, $i \in \{H, L\}$. This is a discrete version of the MLRP assumption. Thus, receiving signal s_L is "good news" for the agent. The signal can be interpreted as a measure of the agent's self-confidence, which determines his motivation to do the task. Note that the principal does not observe the agent's private signal.

A situation where the principal is better informed about the difficulty of the task is not exceptional: it arises whenever the task is new to the agent, whereas the principal has observed other agents working on similar tasks before. The principal may be, for instance, an experienced manager, a teacher or a parent, while the agent is a young employee, a student or a child. In this model a bonus, promised by the principal, affects the agent's motivation via two channels. First, it directly increases the agent's incentives to exert high effort by providing a monetary compensation. Second, because it is offered by an *informed principal*, it potentially affects the agent's beliefs about the difficulty of the task.

Before describing the equilibrium, we emphasize that we present a restricted version of Bénabou and Tirole's model. While our version captures its essential features, the original model is more general and has a much broader set of applications. In particular, the principal may be better informed not only about characteristics of the task, but also about the agent's personal qualities. Although we have restricted the set of feasible bonuses, the main results of Bénabou and Tirole (2003) (their Proposition 1, page 497) still hold:[4]

Proposition 1 *(i) Rewards are positive short-term reinforcers: if both bonuses $b = 0$ and $b = \overline{b}$ are given with positive probability in equilibrium, then the probability that*

[4]Bénabou and Tirole's (2003) proof applies almost verbatim despite the modifications in the model.

the agent exerts effort after $b = \overline{b}$ is higher than after $b = 0$.

(ii) Rewards are bad news: when the task is easy, the principal offers a (weakly) lower bonus: if b_H and b_L are bonuses given with positive probability when costs are high ($c = c_H$) and low ($c = c_L$) respectively, then $b_H \geq b_L$.

(iii) Rewards undermine the agent's assessment of the task's attractiveness: for any $s, s' \in \{s_H, s_L\} : E[c|b = \overline{b}, s] > E[c|b = 0, s']$.

The first claim is straightforward: promising more money for less work would be clearly suboptimal. The second claim relies heavily on the two-sided asymmetric information: the principal is privately informed about the costs of effort, while the agent privately observes the signal about the costs of effort. When the costs are low, the agent is more optimistic on average. Hence, it is cheaper for the principal to rely on his intrinsic motivation and not provide additional incentives. While the presence of two-sided asymmetric information complicates the model, it is an indispensable ingredient. Finally, the third claim captures the essential idea that rewards bring bad news; it follows immediately from the second part of the Proposition.

To make the model nontrivial, we impose several restrictions on parameters. First, we assume that $\Delta P > \overline{b}$; otherwise, the principal would never find it worthwhile to offer a bonus. Moreover, for the agent's decision problem to be non-trivial, we assume that were the agent to know the cost of effort, he would exert effort without a bonus if costs are low but not if costs are high: $c_L < \Delta A < c_H$. Exerting effort without any bonus can be thought of as reflecting the intrinsic motivation. Finally, we assume that the bonus is sufficiently high to make effort attractive even if costs are high: $\overline{b} + \Delta A > c_H$. Under these assumptions, there are two possible types of Perfect Bayesian Nash Equilibria that satisfy the "D1 refinement" (Cho and Kreps (1987)). The first one is a pooling equilibrium in which the principal never gives a bonus. The second type is the more interesting partially separating equilibrium in which the principal never gives the bonus if cost of effort is low, and randomizes between the bonus and no bonus when the cost of effort is high.[5]

[5]For this game, the Intuitive Criterion (Cho and Kreps (1987)) is too weak to eliminate equilibria supported by beliefs that do not seem very plausible. For instance, there may be a pooling equilibrium in which the principal offers a bonus under any costs, sustained by beliefs by the agent

In the experiment we implemented parameters under which the equilibrium outcome is unique and it is partially separating, so that receiving a bonus is informative about the cost of effort. These parameter values are summarized in Table 1. Since effort and the bonus are binary decisions, from here on we simply say that the choice is between effort and no effort, and a bonus or no bonus. It is straightforward to verify that the equilibrium outcome is as follows (see the Appendix for a proof):

Proposition 2 *Given the set of parameters in Table 1, in the unique Perfect Bayesian equilibrium outcome of the game satisfying D1:*

- *The principal offers no bonus if costs are low ($c = c_L$) and randomizes between no bonus and bonus $\overline{b}$ if costs are high ($c = c_H$).*
- *The agent exerts effort if he is promised bonus $\overline{b}$ and/or if receives a good signal; if he obtains a bad signal and is promised no bonus he randomizes between high effort and no effort.*[6]

The model itself does not explicitly take into account social preferences. In the experiment, however, we also allow the principal to provide an up-front fixed wage that is independent of success. This wage may be used as an additional channel to adjust differences in payoffs between the players. Even though some additional Perfect Bayesian Equilibria exist with the fixed wage option, in the Appendix we prove that none of these additional equilibria satisfy the D1 criterion when the agent's private signal is sufficiently precise (i.e., $r > \overline{b}/\Delta P$). The implemented parameters satisfy this condition, so no strictly positive fixed wage is used in equilibrium and Proposition 2 still holds.

2.2 Uninformed Principal

In a control treatment of the experiment we analyze the same model, but assume that the principal does not observe the difficulty of the project when she sets bonus

that no bonus means high costs.

[6]More precisely, under our set of parameters, the principal randomizes between no bonus and bonus $\overline{b}$ with probabilities 1/3 and 2/3 when costs are high; the agent randomizes between high and low effort with probabilities 1/9 and 8/9 after getting no bonus and low signal.

b. Let bonuses b_L, b_H be determined by:

$$\begin{aligned} b_L &= \max\{0, E[c|s = c_L] - \triangle A\}; \\ b_H &= \max\{0, E[c|s = c_H] - \triangle A\}. \end{aligned}$$

Then, the agent's best response is to exert effort if he is offered bonus $b \geqslant b_H$, or if offered a bonus $b \geqslant b_L$ and he received signal s_L. Under our parametrization $b_L = 0$ and $b_H = 7.5$. In the unique Perfect Bayesian equilibrium outcome, the (uninformed) principal offers no bonus, and the agent chooses $e = 1$ if gets a good signal and $e = 0$ otherwise.

3 Experimental set-up and hypotheses

3.1 Design

The experiment implements the model described in the previous section, with parameters as summarized in Table 1. We first describe the main treatment ("informed condition"). In every round, two players are anonymously matched to each other, one in the role of principal, the other in the role of agent. There are two stages. In the first stage, the principal observes the difficulty of the project ($c_L = 15$ or $c_H = 45$) and then specifies the bonus $b \in \{0, 20\}$ and the fixed wage $w \in \{0, 5, 10\}$ for the agent. The fixed wage is paid to the agent irrespective of the agent's choices, while the bonus is paid only if the agent chooses the high level of effort.

In the second stage, the agent (who so far only knows that $c = 15$ or $c = 45$ with equal probabilities) observes the bonus and the wage offered by the principal, and acquires the private signal about the difficulty of the project (which is correct with probability 3/4). Then, he chooses whether or not to exert effort on this *joint project*, $e_J \in \{0, 1\}$. High effort by the agent increases the payoff for both players by $\Delta P = \Delta A = 30$.

A key feature of the design is to introduce the second, *own project* for the agent. The principal derives no benefit from the agent's own project and, therefore, the bonus applies only to the joint project. In all other respects, the two projects are identical; in particular, their cost realizations are perfectly correlated and the agent

receives a single informative signal that applies equally to both projects he is facing. The agent chooses the effort level $e_I \in \{0, 1\}$ that he wants to apply to his own project simultaneously with his choice of e_J. Since the agent receives no bonus for his own project, the bonus cannot be a direct motivator in this case. However, insofar as the agent infers any information from the bonus, this inference will have an equal impact on his effort level in both tasks.

This feature of the experiment allows us to distill the informational aspects of rewards as perceived by the agent from the direct incentive effects. Alternatively, we could have asked the agent to report his beliefs about the costs. However, asking for beliefs would have made it salient that we expect adjustments in beliefs depending on the bonus, prompting participants to think more consciously of this. Our approach is an attempt to minimize this potential problem.

As in many experiments (e.g., Fehr et al. (1993)), the task is one of "stated effort" rather than "real effort." While a real effort task adds realism, an important advantage of using a stated task is that we know precisely the disutility of effort function and can therefore compute the optimal choices in equilibrium. This way, we can control any differences in the ability and/or cost of effort separately from personal characteristics such as risk aversion that may otherwise be correlated with the cost of effort (Charness and Kuhn (2011)). In our context it is particularly important to have precise information about the the exact structure of the two-sided private information and exclude any interference with private information about ability and cost of effort, since we focus on the informational aspects of rewards. Although there is only very limited evidence on this issue, Brüggen and Strobel (2007) provide some evidence that a stated effort task yields qualitatively similar results as a real effort task, and is an appropriate way of implementing an effort decision.

Besides the main treatment we had a control treatment ("uninformed condition"). In the control treatment the principal was not given any private information on the costs. This was common knowledge to the players. By comparing the agents' reaction to bonuses offered by the informed and uninformed principals we have a robustness check to determine the extent to which the agents' reaction to bonuses is explained

by the principals' access to private information about the task.

Finally, in 6 of the 8 sessions, we added a third stage where we measure several dimensions of social preferences. We used this as an extra robustness check to ensure that the behavior we find is not due to other-regarding preferences. For this, we implemented a design based on Cox (2004), with between-subject procedures being replaced by within-subject ones. First, participants were matched in pairs and played a standard trust game. The sender was endowed with 20 points and could send any multiple of five to the receiver (denoted by st, for "sent in trust game"). The amount sent was tripled, and the receiver then decided how much to return (rt, "return in trust game"). Every participant played this game in both roles, using the strategy method for receivers (i.e., asking asking about their reaction to all possible actions by the sender). The main reason for using the strategy method was avoiding emotional spillovers to subsequent periods rather than generating more data. In the third round, every participant played the game once more as a sender, but this time without an option for receivers to return any amount (sd, for "sent in dictator game"). Finally, each participant played once more as a receiver, but now with the amount received being randomly determined by the computer rather than being selected by the matched sender (rr, "returned if amount random"). The computer-generated amount was subtracted from the matched sender's account. Participants faced different partners in different periods.

The purpose of this design is to have a multi-dimensional measure of social attitudes. Based on the data collected we constructed four variables reflecting social preferences. *Altruism* is defined as the fraction out of the endowment sent to the receiver in the dictator game ($sd/20$). The difference in fraction sent between the dictator game and the gift exchange game is used as a proxy for *trust* ($st/20 - sd/20$). We define *fairness* as the fraction of the amount received that is returned to the sender when the amount received was determined randomly ($rr/received$; averaged over the possible positive amounts received). The difference in fraction returned between this treatment and the treatment where the amount received was determined by the sender is defined as the degree of *reciprocity* ($rt/received - rr/received$). We clas-

sify participants that are above median on these measures as "*Altruist*," "*Trusting*," "*Fair*," and "*Reciprocal*."

3.2 Procedures

We ran 8 sessions with 156 participants in total. The number of participants in each session varied between 18 and 24, depending on show up. In four of the sessions we formed independent subgroups with at least 10 subjects in every group to increase the number of independent observations. This gives us a total of 12 independent groups. Participants played a total of 32 rounds of the game, of which 20 rounds in the informed condition, and 12 the uninformed condition.[7] We let them play more rounds in the informed condition because of its relative complexity relative to the uninformed condition. Half of the groups started in the informed condition (I-U groups), the other half started in the uninformed condition (U-I groups). Participants were rematched after every round, to approximate the one-shot nature of the game. Group sizes were too small to ensure that participants never met more than once, but the matching was anonymous and we explained to them that no participant would ever meet the same participant more than once within cycles of 5 consecutive rounds. All players switched roles at certain points, so that they played half of the time as principal and half of the time as agents. Such role switching is commonly used in signaling games to facilitate learning (see, e.g., Brandts and Holt (1992), Cooper and Kagel (2005) and Kübler et al. (2008)). At the end of every round, players observe the cost of the project and payoffs for both players.

The instructions explaining the game were framed in terms of a labor market, using terminology such as principal, worker, wage, and bonus, etc.[8] We explained

[7] In two of the sessions we had a technical problem. In one of these sessions we had to restart the computers after seven rounds in the main treatment. We dropped four participants from the data who could not continue after the interruption and did not finish the entire session. In the other session, we have missing observations for 24 participants for the last eight rounds in the main treatment. We decided to keep these observations, but there are no essential changes in our estimates if they are removed from the analysis. In both cases, all participants completed all rounds of the control treatment.

[8] Cooper (2003) shows that a meaningful context can accelerate learning in experiments with signaling games; Cooper et al. (1999) show that the impact of the context depends crucially on the audience (students vs. managers).

participants that the task of the principal was to determine the bonus and wage, and that of the agent to choose an effort level. We conjecture that most people associate a bonus with something positive. If so, they are, if anything, *less* likely to infer negative information from a bonus than if we would use more neutral terminology, giving a more stringent test of the hypothesis.

The experiment was computerized using Z-tree (Fischbacher (2007)). Sessions took place in 2009-2011 at two Russian universities (NES and ANE, Moscow). Participants were paid for their decisions in every round, with earnings averaging 370 Rubles (approximately $13). Participants in the role of the agent were paid for only one of the two projects determined randomly, to avoid risk hedging behavior (see Blanco et al. (2010)). Sessions lasted for about 90 minutes. All participants were economics students with no or little training in game theory or behavioral economics. A translation of the instructions is included in the Appendix (the original is in Russian).

3.3 Hypotheses

Based on the propositions in the previous section, we formulate the three main hypotheses.

Hypothesis 1 *An informed principal is more likely to offer a bonus when she observes a high level of costs, so that the bonus embodies bad news.*

The first hypothesis implies that a promise of a bonus brings bad news about task difficulty. The second hypothesis stipulates that the positive direct incentive effect of the bonus outweighs this negative information:

Hypothesis 2 *A bonus increases effort by the agent in the joint project.*

The third hypothesis states that the negative information, contained in the bonus, is correctly inferred by the agent and reduces his intrinsic motivation.

Hypothesis 3 *With an informed principal, the agent infers bad news from a bonus and reduces effort in his own project.*

4 Results

4.1 Main treatment

We first discuss the results from the main treatment, and postpone the discussion of the control treatment (uninformed condition) and social preferences to the next subsections.

To be conservative, we always treat the group means as the units of observation when we use nonparametric tests, giving us 12 independent observations for each condition. We did not find any indication of order effects of the conditions (I-U versus U-I groups), so we only report the results of all groups combined.

In the main treatment (informed condition) the principal observes the cost of effort and can adjust the bonus to the cost of effort. We first verify that the bonus is informative about the level of costs, which is a crucial part of the experiment. Figure 1 shows the results. It is indeed the case that the level of the bonus is very informative. When costs are high, the principal gives the bonus 80% of the time, compared to only 32% when the costs are low, and this difference is significant (Wilcoxon signed rank test, $Z = 3.1$, $p = .002$, two-tailed test). This shows that rewards are informative about the cost level.

Table 2 shows the marginal estimates of a probit model with standard errors clustered at the group level.[9] Column 1 shows that if costs are high, the likelihood that a bonus is given increases by 48 percentage points. In column 2, we control for the social preferences measures. Possibly, the relatively fair-minded principals are more likely to give a bonus, in which case the bonus also becomes informative about the fairness of the principal. We do not find any significant effects of the social preferences variables on the likelihood of giving a bonus. The effect of high costs is by far the best predictor of a bonus. In section 4.3 we will discuss potential interaction effects with social preferences. When we only consider the first or last 10 rounds (columns 3 and 4), we see that the coefficient of high costs becomes somewhat larger

[9] We also estimated all specifications using a linear probability model with random or fixed effects at the group level, and a probit model with group random effects. All specifications give very similar results. In particular, the size and significance of our main variable of interest (the impact of a bonus on effort) is robust across different specifications.

in the last 10 rounds, but is already large in the first ten rounds.

Result 1 *A bonus is very informative about the level of costs in the informed condition. High costs increase the likelihood of a bonus by around 50 percentage points.*

This result confirms hypothesis 1.

Before turning to the response by the agents, it is also worthwhile to examine the fixed wages offered by the principals. The vast majority of principals gives a zero fixed wage. This is largely independent of the observed costs. A positive wage is given 23% and 18% of the time when costs are respectively low and high. Figure 2 shows that the principal is only a bit more likely to offer a positive fixed wage when she offers no bonus, and the distribution of fixed wages is very similar after observing high or low costs. Thus, the up-front fixed wage is not very informative about the observed cost level by the principal. The estimates from Table 2 are also essentially unchanged if we analyze the bonus and wage decisions simultaneously in a multivariate probit model (not reported).

We now turn to the behavior of agents. Before we study the impact of a bonus on effort in the own project, we examine the impact on effort in the joint project. In equilibrium, the size of the reward should offset any negative information effects, and have a positive impact on effort in the joint project. Recall also that the agent receives an informative private signal about the cost of effort, giving an indication that costs are low or high. Because the agent's reaction to a bonus can differ depending on the signal, we report results for each signal. We will refer to the private signal of low costs as "*good signal*" as this is positive news for the agent.

A bonus is indeed very effective in stimulating effort in the joint project. When agents receive no bonus, 21% (after a bad private signal) and 60% (after a good private signal) of the agents exert effort. After receiving a bonus, 92% of the agents exert effort in the joint project after each signal. The two most left bars in Figure 3 show the increase in effort split by signal for the main treatment. The difference in effort between a bonus and no bonus is significant for each private signal (in both cases $Z = 3.1$, $p = .002$, signed rank test).

Table 3 reports marginal effects of Probit estimations and confirms the results from the nonparametric tests. We include an interaction term for bonus and good signal since, as mentioned before, the effect of a bonus is expected to be different depending on the signal.[10] Columns 1 and 2 report the marginal effects on effort in the joint project.[11] The effect of receiving a bonus is large and significant whether or not controlling for gender and the social preferences measures (in the next section we discuss some interaction effects with social preferences). As can be seen from the interaction term, the effect of a bonus on effort is smaller after receiving a good signal, because effort is already relatively high in that case even with no bonus. But also in that case the bonus has a significant and large effect on effort. If we estimate these specifications separately for the first set of ten rounds and the second (last) set of ten rounds, we find that the effect of a bonus on effort in the joint project is highly significant in both cases (not reported).

Thus, we can confirm hypothesis 2 for the informed condition.

Result 2 *With an informed principal, a bonus increases effort in the joint project.*

The next question is whether or not rewards are perceived correctly as informative by the agent. We can investigate this by looking at effort in the own project. The bonus offered by the principal does not apply to the own project, so the only reason why a bonus might have an impact is that the agent infers informational content from it. If so, a bonus should reduce effort. Figure 4 shows the increase in effort between a bonus and no bonus. Following a bonus, effort in the own project is 19 percentage points lower after each signal (good and bad signal), and both these differences are significant ($Z = 3.1$, $p = .002$).

Columns 3 and 4 of Table 3 report the marginal effects on effort in the own project. Receiving a bonus substantially reduces the likelihood of exerting effort on the own project. After controlling for social preferences, the coefficient is -19.5 percentage

[10] Here and elsewhere, when we report marginal effects of a Probit regression, the coefficient and standard error of the interaction term are corrected to account for the nonlinear nature of the model, see Ai and Norton (2003) and Norton et al. (2004).

[11] The number of observations is lower if we control for social preferences since we did not collect this information in all sessions. We also have missing information on gender for 7 subjects.

points after receiving a bad signal, and reduced further by another 6.2 percentage points after a good signal. The coefficient is naturally somewhat smaller (in absolute terms) after a bad signal, because the effort is already relatively low in that case.[12] Thus, a bonus is perceived as bad news.

Since it is counter-intuitive that a bonus is bad news, one may expect that participants show some learning over the course of the experiment. Figure 5 plots for every round the mean difference in effort in the own project between a bonus and no bonus (using a 3-round moving average to smooth out some of the variation). Inspection of the figure reveals that there is a clear downward trend in the difference in effort. Columns 5 and 6 of Table 3 show indeed that in the first 10 rounds the effect is mostly absent after a bad signal but already present after a good signal (the coefficients of bonus and bonus X goodsignal are jointly significantly different from zero, $p = .023$). In the last 10 rounds, the negative effect becomes very strong: after receiving a bonus, effort in the own project decreases by 34 percentage points.

Thus, while a bonus is by itself a good motivator, and agents respond positively to a bonus in the joint project, agents also correctly infer that a bonus conveys bad news about costs in the informed condition, and reduce investments in their own project. This effect is particularly strong in the last 10 rounds. We can therefore confirm hypothesis 3.

Result 3 *Agents correctly infer bad news from a bonus in the informed condition, leading them to reduce effort in the own project.*

Overall, our results so far provide clear support for the model.[13] Participants in the role of principal use rewards to stimulate the agents, and agents respond to these rewards as expected, including the correct inference of information.

[12] That the interaction effect of bonus and good signal is stronger for the joint project than for the own project is consistent with the theoretical predictions. The theory predicts a larger effect of a bonus on effort in the joint project after receiving a bad signal. Since the effort on the own project is not restricted after receiving a bad signal and no bonus (the agent is indifferent between effort and no effort in that case), the interaction effect need not be there.

[13] The result that high-powered incentives in one dimension crowd out effort in the other is also consistent with the seminal model on multi-tasking by Holmstrom and Milgrom (1991). However, here the mechanism is purely informational, while in that model crowding out occurs because of convex costs and substitutability of effort on different dimensions.

4.2 Control treatment

As a further robustness check of the model, we also implemented the control treatment in which the principal is not informed. In this uninformed condition, we still expect that agents respond positively to rewards in the joint project. However, since rewards are not informative, we do not expect that effort in the own project varies with the bonus.

In the joint project, we find indeed that effort is higher after receiving a bonus (see the right two bars in Figure 3). Both after a good and a bad signal, this effect is significant (signed rank test, $Z = 3.1$, $p = .002$). This is also confirmed in the regression analysis shown in columns 1 and 2 of Table 4. Turning to the own project, the effort level of agents does not vary much with respect to the bonus, as expected, and any differences are not significant using nonparametric tests (see Figure 4). After receiving a good signal, effort is 5 percentage points lower following a bonus ($Z = 1.1$, $p = .267$). After receiving a bad signal, effort is 1 percentage point lower after a bonus ($Z = .12$, $p = .906$). The regressions show that effort is not significantly different after a bonus when the signal is bad (columns 3 and 4 of Table 4). After a good signal, the total effect of a bonus is significantly negative, which is somewhat surprising. However, the coefficient is modest in size (around minus 8 percentage points), and if we estimate the model separately for each signal the coefficient of bonus is always small and not consistent in sign.[14]

So far, the reported results are based on all sessions combined independent of the treatment order. Since the only difference between the main and control treatment is whether or not the principal is informed, going from one treatment to the next may have made it salient participants that there are possible information effects. However, if we repeat the above analysis for the subset of participants in their first treatment, the results are very similar. In particular, the reported means in Figures 1, 3, and 4, are very similar. A bonus is significantly more likely if costs are high, the effort in the joint project is significantly higher after a bonus in each treatment and for each

[14] As in the main treatment, the theory predicts a negative interaction effect of bonus and good signal for the joint project, and no interaction effect for the own project, which is what we find.

signal, while the effort in the own project is significantly lower after a bonus in the main treatment (all p-values in these cases are smaller than .03). Effort in the own project is not significantly lower after a bonus in the control treatment ($p = 0.75$ after a bad signal, $p = .29$ after a good signal).

4.3 The role of social preferences and gender

We now investigate in some more detail the role of social preferences. Many studies have shown that people care about the distribution of payoffs and the intentions of others. There is little reason to suspect that social preferences are driving our key result that a bonus is perceived as bad news. In particular, the effort decision for the own project does not affect payoffs for the other participant, so there is no reason to expect that the negative effect of a bonus in the own project is driven by social preferences rather than reflecting informational effects. Furthermore, if principals are concerned about inequalities in payoffs resulting from not offering a bonus, they could partially address this by adjusting the fixed wage. Nevertheless, we believe it is interesting to examine the extent to which the response to a bonus in the joint project is driven by social preferences.

In the role of principal, we find very little evidence that social preferences determine the level of the bonus in any substantial way. In table 2, we already showed that the costs are the most important determinant of the bonus. None of the measures of social preferences has a significant impact. We also did not find evidence of any substantial interaction effects. That is, the estimated coefficient of high costs is broadly similar if model (1) of table 2 is estimated separately for the subsets of participants who are above and below the median for each of the measures of social preferences. We also find little evidence that social preferences play a role in the effort decision in the joint project. Most coefficients related to social preferences are insignificant and relatively small. "Fair" agents tend to exert somewhat less effort in the joint project and "Altruists" a bit more (column 2 of Tables 3).

Arguably the most interesting finding with respect to social preferences concerns the response to different levels of the fixed wage. Figure 6 plots the mean effort in the

joint project for the two conditions, separately for no bonus and bonus (solid lines). In the main treatment (informed condition, left panel), the effort does not vary much with the fixed wage. In the control treatment (uninformed condition, right panel), we find a U-shaped pattern: the mean effort is lower after a fixed wage of 5 than after no wage, but then increases again when the fixed wage is 10. The regressions also show a negative effect of offering a fixed wage of 5 in the uninformed condition (Table 4, columns 1 and 2) but not in the informed condition (Table 3, columns 1 and 2).[15] Possibly, participants think that a small fixed wage is more of an insult if the principal is uninformed, because in that case the principal is not offering this as a compensation for high costs. This, however, is speculation.

This wage effect is reminiscent of the "small payment" effect found in other experiments, such as in Gneezy and Rustichini (2000b), who also find that motivation is lower for small payments than under no compensation at all, but increases for higher payments. However, they find this using piece rates, while in our case we find that pattern with respect to a fixed wage. We also find that the U-shaped pattern is more pronounced for participants with a level of reciprocity that is above the median (the dashed lines in Figure 5). It is possible that agents interpret a fixed wage of 5 as coming from an unfair principal. An ordered probit analysis indicates that "fair" principals are somewhat more likely to offer a positive fixed wage, while none of the other social preferences has an effect.[16] The effect is somewhat larger in the main treatment with an informed principal (roughly 12 percentage points) than in the control treatment (roughly 5 percentage points), but there is not much difference in the higher likelihood of offering a wage of 5 or 10 by fair principals. Perhaps principals only rarely offer a fixed wage of 5 because they realize that this has an averse effect on effort.[17] In any case, because of the rare occurrences, all these results should be

[15] Estimates from a linear probability model with group fixed effects deviate from those in models (1) and (2) in Table 4. The negative effect of wage 5 is smaller, and the wage 10 coefficient is smaller and not significant in that case.

[16] We do not report these results for sake of brevity, but a more detailed analyisis is available upon request.

[17] We do not have a reliable number of observations to test significance using the means of groups as independent observations. If we treat every choice of a subject as an independent observation, the difference between 0 and 5 is significant after a bonus ($p = .000$) and at the margin of significance after no bonus ($p = .125$). The difference between 5 and 10 is only significant after no bonus

taken with some caution.

While we control for gender in all regressions, we find almost no significant gender effects. More precisely, the only significant gender effect is that effort is higher for women in the own project of the main treatment, but this disappears in later rounds (column 6 of Table 3). This result may nevertheless be important for the sake of avoiding a publication or reporting bias against finding null results (no gender difference), as argued by Croson and Gneezy (2009).

5 Discussion

Our experiment shows that when the principal is better informed about characteristics of the task than the agent, rewards have hidden costs as predicted by Bénabou and Tirole (2003). The principal is more likely to offer a bonus when she knows the task to be difficult, and this is correctly perceived by the agent.

The experimental design allows us to isolate the informational effects of rewards. Of course, by no means does this imply that other factors such as social preferences are not important. As discussed in the introduction, a large experimental literature shows that rewards may have a strong negative impact on motivation even when the principal does not have superior information about the task, which is a key assumption in our set-up. The main channel in that case is the impact of rewards on fairness, reciprocity and trust related concerns. We view our paper as an important complement to that literature, showing that the interaction of extrinsic incentives and intrinsic motivation is a multifaceted phenomenon that cannot be reduced to a single idea or theory.[18] Investigation of the interaction between pure informational and fairness-related effects seems to be an important topic for future research, both

($p = .025$). The difference between 0 and 10 is significant in both cases ($p = .006$ after no bonus, $p = .024$ after bonus). None of the differences is significant in the informed condition (all χ^2 tests).

[18] For instance, the model by Bénabou and Tirole (2006) demonstrates that rewards and punishments can have a negative impact on prosocial behavior because they create doubts about the true motives of altruistic behavior and thus reduce the importance of concerns for social and self-respect. Janssen and Mendys-Kamphorst (2004) pays special attention to the dynamics of prosocial behavior when extrinsic incentives for such behavior change. Seabright (2009) analyzes prosocial behavior in a signaling and screening context taking into account assortative matching between agents it induces. Ellingsen and Johannesson (2008) explore how the principal's choice of incentive scheme, being informative about her character, affects the agent's desire to seek the principal's esteem.

theoretical and experimental.

A natural extension would be to conduct an experiment with a real effort task, to test the external validity of our findings. The challenge will be to implement the required two-sided information asymmetry in a controlled manner. Another robustness check would be to investigate the impact of role switching. On the one hand, we feel that this element of design helps participants to understand the features of that game more quickly. In reality, people have more time to learn than we give them in the lab. On the other hand, many people are always on the same side of the relationship and may not have an opportunity or incentive to take another perspective. For instance, some people will never hold a managerial job, and such people may fail to understand the exact motives behind the choice of rewards by the employer.

Since in real life the hidden costs we explore may have only a delayed impact, an important venue for further experimental research is the study of repeated relationships. A model in Suvorov (2003) shows that in this case rewards become "addictive" if the agent's opportunities to independently acquire information about the task are limited. Two new strategic effects arise in the model: the agent tries to appear unmotivated to convince the principal to give a high bonus in the future, and the principal is concerned about promising a bonus and thus creating "addiction to rewards".

Finally, we would like to emphasize that the experimental research of the information transmission via rewards need not be restricted to an investigation of a negative impact. For instance, the model in Suvorov and van de Ven (2009) shows that non-contractible ex post rewards can occur even in a finitely-repeated relationship if the principal has superior information about the agent's interim performance. It shows that rewards may also have informational content in that case, but the information becomes good rather than bad news. Such discretionary rewards signal that the principal appreciates previous efforts and has high expectations about future achievements, thus giving a boost to the agent's motivation in the remaining periods.

6 Appendix A: Proofs

Proof of Proposition 2.

From part (ii) of Proposition 1 it follows that there are five possible types of equilibria: two types of pooling equilibria (the principal always giving no bonus or always giving bonus $\overline{b}$), a separating equilibrium (no bonus when costs are low, $\overline{b}$ when costs are high) and two types of hybrid (partially separating) equilibria. In hybrid equilibria of the first type, no bonus is offered when the project is easy, and the principal randomizes between no bonus and bonus $\overline{b}$ when it is difficult. In the second type of hybrid equilibria, bonus $\overline{b}$ is offered when the project is difficult, and the principal randomizes between no bonus and bonus $\overline{b}$ when it is easy.

Note first that the separating equilibrium and the second type of partially separating (hybrid) equilibria cannot occur under our assumptions. In such equilibria the principal would always prefer to deviate and give no bonus as this induces the agent to exert high effort.

Moreover, a pooling equilibrium in which the principal offers a strictly positive bonus is eliminated by D1 (or by NWBR, which is equivalent to D1 in the current game).[19] Suppose, by contradiction that the equilibrium is pooling with bonus $\overline{b}$ always being offered. The agent always exerts effort in this case. Let the agent's response to the out-of-equilibrium bonus $b = 0$ be such that he chooses $e = 1$ with probabilities μ_H and μ_L when his signal is s_H and s_L respectively. Consider the response by the agent which would make the principal indifferent between offering $b = \overline{b}$ and deviating to $b = 0$ when costs are high ($c = c_H$). Then the principal would strictly gain from deviation to $b = 0$ when costs are low ($c = c_L$) whenever $(2r-1)\Delta P(\mu_L - \mu_H) > 0$. It is straightforward to show that the latter condition holds, given that the principal's indifference implies $(\mu_H, \mu_L) \neq (1,1), (0,0)$, and given that it must be the case that $\mu_H < \mu_L$ for the agent's strategy to be a (mixed) best response for some (out-of-equilibrium) beliefs. The NWBR criterion then stipulates that the agent should assign probability 0 to $c = c_H$ after observing $b = 0$, giving

[19] For a general definition of D1 and NWBR refinements we refer the reader to Cho and Kreps (1987); Cho and Sobel (1990) prove that they are equivalent in montonic games. In our model NWBR is defined as follows. Consider the agent's reaction to an out-of-equilibrium offer b' that is (a) a best response under some beliefs and (b) makes the principal indifferent between sticking to an equilibrium action and deviating to b' when the cost is $c = c_i$. For an equilibrium to satisfy NWBR, out-of equilibrium beliefs must assign probability 0 to the value of cost c_i if the principal strictly gains from the deviation to b' under this agent's reaction if the cost $c = c_j \neq c_i$.

incentives for the principal to deviate, which upsets the equilibrium (see Cho and Kreps (1987)).

It is straightforward to check that under the chosen values of parameters the strategies specified in Proposition 2 indeed form a hybrid equilibrium, while the pooling equilibrium with no bonus does not occur. In the pooling equilibrium with $b = 0$, the agent works after signal s_L but not after signal s_H. Thus, if costs are high, the principal expects the agent to exert effort with probability $1 - r$ (the likelihood that the signal is incorrect). The principal would prefer deviating to $\overline{b}$ (inducing the agent to exert effort for any signal) if $\overline{b} < r\Delta P$, which is the case under our parameters.

Proof that $w = 0$ in any PBE satisfying D1 if the agent's signal is precise enough ($r > \overline{b}/\Delta P$).

Note first that a contract $(w, \overline{b})$ with $w > 0$ cannot be offered with a positive probability in any PBE: with this bonus the agent always exerts effort for any beliefs about the costs, so this contract is strictly dominated by the contract $(0, \overline{b})$.

Next, let us show that if contract (w_H, b_H) is offered with some probability when $c = c_H$ and (w_L, b_L) is offered with some probability when $c = c_L$, then $b_L \leq b_H$. Assume by contrast that $b_L > b_H$, i.e. $b_L = \overline{b}$, $b_H = 0$. Hence, $w_L = 0$. Let the agent, when offered $(w_H, 0)$, choose $e = 1$ with probabilities μ_H and μ_L if his signal is s_H and s_L respectively. For the agent's effort choice to be a best response under some beliefs, it must be that $\mu_H \leq \mu_L$ and $\mu_L = 1$ if $\mu_H > 0$. By a simple revealed preference argument, the principal must weakly prefer $(0, \overline{b})$ to $(w_H, 0)$ when costs are low, and weakly prefer $(w_H, 0)$ to $(0, \overline{b})$ when costs are high, so that:

$$\begin{aligned} \Delta P - \overline{b} &\geq \Delta P(r\mu_L + (1-r)\mu_H) - w_H, \\ \Delta P(r\mu_H + (1-r)\mu_L) - w_H &\geq \Delta P - \overline{b}. \end{aligned}$$

Since we must have $\mu_H \geq \mu_L$, these inequalities imply that $\mu_H = \mu_L$. If $\mu_H = \mu_L = 0$, the principal would prefer to deviate to $(w, b) = (0, \overline{b})$ if $c = c_H$. If, alternatively, $\mu_H = \mu_L = 1$, then (from the same revealed preference argument) it follows that $w_H = \overline{b}$. Then, the principal gets $\Delta P - \overline{b}$ in equilibrium. D1 implies (see the argument

in the previous proof) that beliefs after contract $w = 0, b = 0$ should be that $c = c_L$, which destroys the equilibrium.

Similar arguments as above imply that $(w, b) = (0, 0)$ should be offered on the equilibrium path with a positive probability in both cases, i.e., if $c = c_L$ and if $c = c_H$, as is easy to verify.

Assume now that contract $(w, 0)$ with $w > 0$ is offered with a positive probability. If this contract were offered only in case $c = c_H$, the agent would exert no effort, and the principal would deviate to $(0, 0)$. Assume now contract $(w, 0)$ is offered in case $c = c_L$ only. Then the agent is sure to exert effort if offered this contract. Denote again by μ_H and μ_L the probabilities that the agent exerts effort if offered contract $(0, 0)$ and his signal is s_H and s_L respectively. Then, since the principal must be indifferent between $(w, 0)$ and $(0, 0)$ when costs are low and weakly prefer $(0, 0)$ to $(w, 0)$ when costs are high, we have:

$$\begin{aligned} \Delta P - w &= \Delta P(r\mu_L + (1-r)\mu_H), \\ \Delta P - w &\leq \Delta P((1-r)\mu_L + r\mu_H). \end{aligned}$$

Since $\mu_H \leq \mu_L$ and $\mu_L = 1$ if $\mu_H > 0$, this implies $\mu_H = \mu_L = 1$ or $\mu_H = \mu_L = 0$. We get a contradiction: the first option violates $a > 0$, the second implies that the principal would want to deviate to $(0, \overline{b})$ under both cost realizations.

Hence, the principal should offer both contracts $(0, 0)$ and $(w, 0)$ with a positive probability in both cases, $c = c_L$ and $c = c_H$. Denote by $\hat{\mu}_H$ and $\hat{\mu}_L$ the probabilities that the agent exerts effort if offered contract $(w, 0)$ and his signal is s_H and s_L respectively. For the principal to be indifferent between the contracts we must have:

$$\begin{aligned} \Delta P(r\mu_L + (1-r)\mu_H) &= \Delta P(r\hat{\mu}_L + (1-r)\hat{\mu}_H) - w, \\ \Delta P(r\mu_H + (1-r)\mu_L) &= \Delta P(r\hat{\mu}_H + (1-r)\hat{\mu}_L) - w. \end{aligned}$$

This implies $\hat{\mu}_L - \mu_L = \hat{\mu}_H - \mu_H > 0$. This is possible only if $\mu_H = 0, \hat{\mu}_L = 1$ and $0 < \hat{\mu}_H = 1 - \mu_L < 1$. However, if $r > \overline{b}/\Delta P$, then $\Delta P - \overline{b} > \Delta P(1-r)\mu_L$, so that the principal is strictly better off when she offers contract $(0, \overline{b})$ if $c = c_H$ – a contradiction.

7 Appendix B: Instructions [Not for publication].

The following instructions are translated from Russian. These are the instructions for groups that were first in the main treatment and then in the control treatment. The instructions for the reverse treatment are essentially the same and available upon request.

Please read these instructions carefully. You will have a chance to earn a considerable amount of money if you read the instructions carefully. The exact amount depends on your own choices and the choices of other participants. You can collect your earnings immediately after the experiment. All your choices will remain confidential, and nobody else besides the researchers will know how much you earned. **It is prohibited to communicate with other participants during the experiment!** If you violate this rule we will exclude you from the experiment and you will not receive your earnings. All participants in your session receive the same set of instructions. Please raise your hand if you have any questions and one of us will come to you.

The experiment consists of three parts. You first get instructions for the first part. The instructions for the other parts will be handed out later. At the end of the experiment you will be asked to enter your identification number, located on the sheet that you were given when you entered the room. This number will be used to calculate your earnings.

Part 1 instructions. This part describes the general setup.

Two persons, whom we call the Principal and the Worker, are working on a joint project. The Worker may exert a high or a low level of effort by choosing $e_J = 1$ or $e_J = 0$. If the Worker chooses the low level of effort ($e_J = 0$), he bears no costs ($c = 0$). If he chooses the high level of effort ($e_J = 1$), he bears cost c. The size of c depends on the project difficulty and is either a high value of $c_H = 45$ points (for a "difficult" project) or a low value of $c_L = 15$ points (for an "easy" project). Both values of c are equally likely to occur.

If the Worker chooses the low level of effort, the project fails and yields $A_0 = 25$ points to the Worker and $P_0 = 10$ points to the Principal. If instead the Worker chooses the high level of effort, the project succeeds and yields an additional $\Delta A = 30$ to the Worker and $\Delta P = 30$ to the Principal, i.e., in this case they receive $A_1 = 55$ and $P_1 = 40$ respectively.

The Principal is fully informed about the difficulty of the project (the level of c), while the Worker does not know the exact value of c, but obtains a signal s, which can assume one of two values: "c_H] or "c_L]. The signal is correct (equal to the true value of c) with probability 3/4 and incorrect with probability 1/4. Unlike the Worker, the Principal does not observe signal s.

The interaction between the two participants proceeds as follows. The Principal observes the project difficulty c and assigns a bonus b she will pay to the agent in case of a successful project. The bonus can assume either of two values: 0 or 20 points. The Principal also determines the fixed salary a she will pay regardless of whether the project succeeds or fails. The fixed salary can be 0, 5 or 10 points. The Worker then observes the values of a and b chosen by the Principal, as well as signal the s. The Worker then chooses the level of effort e, which determines the success of the project.

The Worker is also involved in his individual project which has the same characteristics as the joint project that is just described. In particular, the cost of effort, still not observable by the agent, is the same as in the joint project. The worker chooses an effort level for this individual project ($e_I = 1$ or $e_I = 0$) in addition to the effort level for the joint project.

The Principal does not derive any payoffs from the Worker's individual project and hence the compensation offered by the Principal to the worker does not apply to the individual project.

The joint project therefore yields a payoff to the Principal that is equal to $U_P = P_0 + (P_1 - P_0 - b)e_J - a$ and a payoff to the Worker equal to $U^J_{AH} = A_0 + (A_1 - A_0 + b - c_H)e_J + a$ if the project turns out to be difficult or $U^J_{AL} = A_0 + (A_1 - A_0 + b - c_L)e_J + a$ if the project turns out to be easy. The individual project yields a payoff to the Worker equal to $U^I_{AH} = A_0 + (A_1 - A_0 - c_H)e_I$ or $U^I_{AL} = A_0 + (A_1 - A_0 - c_L)e_I$, depending on the project difficulty.

In each round the Worker only earns points for one of the two projects (joint or individual), which is determined randomly at the end of the round after both levels of effort e_J and e_I are chosen. The Principal always gets points for the joint project only.

Experimental procedures

The interaction described in the previous section will be repeated for 20 rounds. At the beginning of the experiment all participants are split into two groups of equal size – Principals and Workers. Each participant retains his or her role for 5 rounds, then roles are switched for the next five rounds, etc. If you start as a Principal, then you are a Principal in rounds 1-5 and 11-15 and a Worker in rounds 6-10 and 16-20. Similarly, if you start as a Worker, then you are a Worker in rounds 1-5 and 11-15 and a Principal in rounds 6-10 and 16-20.

You will be rematched to another participant in every round. You will not be able to identify the participant with whom you are matched (and (s)he cannot identify you). Within every five round cycle you will never be matched to the same participant.

If you are a principal, you learn the difficulty of the project in that round. You then will be asked to enter a salary a and bonus b that you assign to the Worker. These values are then translated to the Worker who also observes signal s (which you do not observe at this point). After the Worker chooses an effort level, you will be informed about the outcome of the project, as well as about the signal s received by the Worker. Depending on the success of the project you will be credited with either $P_0 = 10$ or $P_1 = 40$ points. The salary that you assigned to the Worker will be subtracted from this. In case of success, the bonus will also be subtracted from your earnings. At the end of the round, you will also learn which level of effort the Worker chose for his or her individual project.

If you are a worker, you observe signal s about the difficulty of the project (your Principal does not observe your signal s at this point), and also the values of the salary a and bonus b assigned by the Principal. You will then be asked to choose your effort level in the joint project e_J and your effort level in your individual project e_I. You will then be informed about the level of project difficulty, the outcome of the project, and also for which of the two projects you earned points in that round.

THROUGHOUT THE FIRST PART OF THE EXPERIMENT THE CONVERSION RATE IS **1 POINT = 0.3 RUBLES**

If you have questions about the first part of the experiment, please ask them now.

Part 2 instructions

The second part of the experiment is very similar to the first part: it will again have four cycles, in which your role will alternate between the Principal and the Worker. Now each cycle will consist of 3 rounds, making a total of 12 rounds. Your payoff will be

determined by the same formulas as before. The only difference is that IN THIS PART THE PRINCIPAL HAS NO INFORMATION ABOUT THE DIFFICULTY OF THE PROJECT. In the beginning of each round the Worker receives signal s about the difficulty of the project. The Principal, as before, can offer a bonus to the Worker, to be paid in case the project is successful, as well as a fixed salary. Based on the signal and the salary and bonus offer, the Worker chooses the levels of effort in the joint and the individual projects. At the end of each round the two participants learn the same information as before.

During each three round cycle you will be matched with different participants. The first cycle starts with the same roles as in part 1 of the experiment.

THROUGHOUT THE SECOND PART OF THE EXPERIMENT THE EXCHANGE RATE IS **1 POINT = 0.3 RUBLES**

If you have questions about the second part of the experiment, please ask them now.

Part 3 instructions

The third part of the experiment differs substantially from the first two: it will consist of four different rounds. What now follows is a description of each round.

THROUGHOUT THE THIRD PART OF THE EXPERIMENT THE EXCHANGE RATE IS **1 POINT = 1 RUBLE**

We first describe the setup that is common to all rounds. Two participants are paired, whom we call the sender and the receiver. The sender gets $S = 20$ points and can send x points to the receiver (x can be 0, 5, 10, 15 or 20 points). The amount sent x is then tripled, so the receiver gets $3x$. The receiver can then return an amount z back to the sender, which can be any amount between 0 and $3x$. The sender then earns the amount $S - x + z$, and the receiver earns $3x - z$.

In the beginning of the first round, you will be matched with two other participants: in the first match you will be in the role of the sender and in the second match you will play the role of the receiver. The two different participants will remain your partners for the first two rounds.

In the first round you will be playing as a sender. You will be endowed with $S = 20$ points and you can decide on the amount x that you like to send to your receiver (x can be 0, 5, 10, 15 or 20 points). This will conclude the first round.

In the second round you will be playing as a receiver and will receive the amount $3x$ from the sender. You will only find out at the end of the experiment how much the receiver has sent to you. You can decide which amount z you would like to return to your sender for every possible amount (s)he may have sent to you (i.e., 0, 5, 10, 15 or 20 points). You will therefore have to enter four numbers (since you cannot return anything if you receive zero points). This will conclude the second round. At the end of the second round you will have earned $R_{12} = S - x + z' + 3x' - z$, where S is your endowment, x and z are your choices in the first and second round, x' is the choice of your sender partner in the first round and z' is the choice of your receiver partner in the second round that corresponds to your choice of x. You will only find out at the end of the experiment how much you have earned in this round.

In the third round you will be playing the role of the sender. In the beginning of the third round you will be matched to a new receiver (and you yourself will also be the receiver matched to some sender). This match will hold for the third round only. You will be endowed with $S = 20$ points and asked which amount x you like to send to your receiver (x can be 0, 5, 10, 15 or 20 points). The receiver will get $3x$. You will only find out at the

end of the experiment how much you have earned. *The only difference with the first round is that the receiver does not have an option to send anything back in this round.* You payoff for the third round will therefore be equal to $R_3 = S - x + 3x'$, where S is the original endowment, x is your choice and x' is the choice of the sender you are matched to. You will only find out at the end of the experiment how much you have earned in this round.

In the fourth round you will be playing as a receiver. You will be endowed with $S = 20$ points. In the beginning of the round you will be matched to a new sender (and you will also be a sender for someone). As in the second round, you will earn $3x$. *The only difference with the second round is that in this round the amount x that is sent to you is randomly chosen by the computer*. It can be 0, 5, 10, 15 or 20 with equal probabilities. The sender is not making the choice of x in this round, but this amount will be subtracted from his or her endowment of $S = 20$ points at the end of the round. You can decide which amount z you like to return to your sender for each possible value of x (0, 5, 10, 15 or 20 points). You will therefore again have to enter four numbers (since you cannot return anything if you receive zero points). In this round you will earn the amount $R_4 = S - x + z' + 3x' - z$, where S is your endowment, x and x' are the amounts chosen by the computer on your and your sender partner's behalf, z and z' your choice and the choice of your receiver partner.

At the end of the third part you will find out how much you have earned in total for all rounds in this part $R_{12} + R_3 + R_4$.

If you have questions about the third part of the experiment, please ask them now.

References

Ai, C. and Norton, E. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1):123–129.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large Stakes and Big Mistakes. *Review of Economic Studies*, 76(2):451–469.

Bem, D. J. (1967). Self-perception: an alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74(3):183–200.

Bénabou, R. and Tirole, J. (2003). Intrinsic and Extrinsic Motivation. *Review of Economic Studies*, 70(3):489–520.

Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *American Economic Review*, 96(5):1652–1678.

Blanco, M., Engelmann, D., Koch, A. K., and Normann, H.-T. (2010). Belief elicitation in experiments: is there a hedging problem? *Experimental Economics*, 13(4):412–438.

Brandts, J. and Holt, C. A. (1992). An Experimental Test of Equilibrium Dominance in Signaling Games. *American Economic Review*, 28(5):1350–1365.

Brüggen, A. and Strobel, M. (2007). Real effort versus chosen effort in experiments. *Economics Letters*, 96(2):232–236.

Charness, G., Cobo-reyes, R., Jiménez, N., and Lacomba, J. A. (2010). The Hidden Costs of Control : Comment. *Mimeo*.

Charness, G. and Kuhn, P. (2011). *Chapter 3, Lab Labor: What Can Labor Economists Learn from the Lab?*, volume 4 of *Handbook of Labor Economics*. Elsevier Inc.

Cho, I. and Sobel, J. (1990). Strategic stability and uniqueness in signaling games. *Journal of Economic Theory*, 50(2):381–413.

Cho, I.-K. and Kreps, D. M. (1987). Signaling Games and Stable Equilibria. *The Quarterly Journal of Economics*, 102(2):179.

Cooper, D. (2003). The impact of meaningful context on strategic play in signaling games. *Journal of Economic Behavior & Organization*, 50(3):311–337.

Cooper, D. J. and Kagel, J. H. (2005). Are two heads better than one? Team versus individual play in signaling games. *The American economic review*, 95(3):477–509.

Cooper, D. J., Kagel, J. H., Lo, W., and Gu, Q. L. (1999). Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers. *American Economic Review*, 89(4):781–804.

Cox, J. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2):260–281.

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

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18(1):105–115.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668.

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior (Perspectives in Social Psychology)*. Plenum Press.

Dickinson, D. and Villeval, M. C. (2004). Does Monitoring Decrease Work Effort ? The Complementarity Between Agency and Crowding-Out Theories. *Mimeo*, 33(0).

Eisenberger, R., Pierce, W. D., and Cameron, J. (1999). Effects of Reward on Intrinsic Motivation: Negative, Neutral, and Positive: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin*, 125(6):691–677.

Ellingsen, T. and Johannesson, M. (2008). Pride and Prejudice: The Human Side of Incentive Theory. *American Economic Review*, 98(3):990–1008.

Falk, A. and Kosfeld, M. (2006). The Hidden Costs of Control. *The American Economic Review*, 96(5):1611–1630.

Fehr, E. and Gächter, S. (2001). Do Incentive Contracts Crowd Out Voluntary Cooperation? *Mimeo*.

Fehr, E., Gachter, S., and Kirchsteiger, G. (1997). Reciprocity as a Contract Enforcement Device: Experimental Evidence. *Econometrica*, 65(4):833 – 860.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation. *The Quarterly Journal of Economics*, 108(2):437–459.

Fehr, E. and List, J. A. (2004). The Hidden Costs and Returns of IncentivesTrust and Trustworthiness Among CEOs. *Journal of the European Economic Association*, 2(5):743–771.

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–40.

Fehr, E. and Schmidt, K. M. (2007). Adding a Stick to the Carrot? The Interaction of Bonuses and Fines. *American Economic Review*, 97(2):177–181.

Fehr, E. A. F. (2002). Psychological foundations of incentives. *European Economic Review*, 46(4-5):687–724.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press, Stanford, CA.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Frey, B. S. (1997). *Not just for the money : an economic theory of personal motivation*. Elgar, Cheltenham.

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

Galbiati, R., Schlag, K., and van der Weele, J. (2009). Can Sanctions Induce Pessimism? An Experiment. *Mimeo*, pages 1–35.

Gneezy, U. and Rustichini, A. (2000a). A Fine Is a Price. *The Journal of Legal Studies*, 29(1):1–29.

Gneezy, U. and Rustichini, A. (2000b). Pay Enough or Don't Pay at All. *Quarterly Journal of Economics*, 115(3):791–810.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics & Organization*, 7.

Janssen, M. C. and Mendys-Kamphorst, E. (2004). The price of a price: on the crowding out and in of social norms. *Journal of Economic Behavior & Organization*, 55(3):377–395.

Kruglanski, A. W., Friedman, I., and Zeevi, G. (1971). The effects of extrinsic incentive on some qualitative aspects of task performance1. *Journal of Personality*, 39(4):606–617.

Kübler, D., Müller, W., and Normann, H.-T. (2008). Job-market signaling and screening: An experimental comparison. *Games and Economic Behavior*, 64(1):219–236.

Lepper, M. R., Greene, D., and Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, 28(1):129–137.

Lepper, M. R., Henderlong, J., and Gingras, I. (1999). Understanding the Effects of Extrinsic Rewards on Intrinsic Motivation - Uses and Abuses of Meta-Analysis: Comment on Deci, Koestner, and Ryan (1999). *Psychological Bulletin*, 125(6):676–669.

Norton, E. C., Wang, H., and Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, 4(2):154 – 167.

Seabright, P. B. (2009). Continuous Preferences and Discontinuous Choices: How Altruists Respond to Incentives. *The B.E. Journal of Theoretical Economics*, 9(1).

Sliwka, D. (2007). Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes. *American Economic Review*, 97(3):999–1012.

Suvorov, A. (2003). Addiction to Rewards. *Mimeo*.

Suvorov, A. and van de Ven, J. (2009). Discretionary rewards as a feedback mechanism. *Games and Economic Behavior*, 67(2):665–681.

Table 1 – Parameter values

Principal		value(s)
Value of project if effort $e = 0$	P_o	10
Value of project if effort $e = 1$	P_1	40
Bonus	b	$\{0, 20\}$
Fixed wage	w	$\{0, 5, 10\}$
Agent		
Value of project if effort $e = 0$	A_o	25
Value project if effort $e = 1$	A_1	55
Likelihood that private signal about costs is correct	r	0.75
Cost of effort if costs are low	c_L	15
Cost of effort if costs are high	c_H	45

Table 2: Bonus, Main treatment

	(1) all rounds	(2) all rounds	(3) rounds 1-10	(4) rounds 11-20
High costs	0.480***	0.502***	0.460***	0.560***
	(0.041)	(0.048)	(0.041)	(0.070)
Female		0.002	-0.043	0.070
		(0.077)	(0.080)	(0.100)
Altruist		-0.007	0.007	-0.028
		(0.053)	(0.060)	(0.075)
Trusting		-0.017	-0.033	0.011
		(0.068)	(0.057)	(0.084)
Fair		0.025	-0.015	0.085
		(0.058)	(0.056)	(0.077)
Reciprocal		0.043	0.046	0.043
		(0.059)	(0.054)	(0.081)
Number of observations	1,461	1,001	547	454
Number of participants	156	110	110	110
Number of groups	12	8	8	8
Pseudo R-squared	0.181	0.203	0.182	0.242

Probit estimates, reporting marginal effects. Robust s.e. clustered at the the group level in parentheses. All specifications include the treatment order as a control variable. *** p<0.01, ** p<0.05, * p<0.1

Table 3: Effort in Main Treatment

	(1)	(2)	(3)	(4)	(5)	(6)
	joint project		own project			
	all rounds	all rounds	all rounds	all rounds	rounds 1-10	rounds 11-20
Bonus	0.636***	0.655***	-0.230***	-0.195***	-0.062	-0.342***
	(0.036)	(0.054)	(0.059)	(0.069)	(0.094)	(0.073)
Good signal	0.304***	0.316***	0.498***	0.480***	0.511***	0.467***
	(0.033)	(0.056)	(0.064)	(0.090)	(0.102)	(0.092)
Bonus X good signal	-0.367***	-0.358***	0.008	-0.062	-0.125	-0.013
	(0.045)	(0.064)	(0.049)	(0.064)	(0.091)	(0.089)
Wage5	0.017	0.047	0.086***	0.070*	0.057	0.100
	(0.044)	(0.047)	(0.031)	(0.038)	(0.054)	(0.062)
Wage10	0.037	0.093*	-0.060	-0.071	-0.020	-0.113
	(0.042)	(0.050)	(0.075)	(0.076)	(0.102)	(0.076)
Female		-0.049		0.106***	0.144***	0.060
		(0.050)		(0.032)	(0.040)	(0.081)
Altruist		0.053**		-0.042	-0.104**	0.042
		(0.023)		(0.036)	(0.050)	(0.062)
Trusting		0.051		-0.010	-0.062**	0.067
		(0.033)		(0.028)	(0.032)	(0.059)
Fair		-0.116***		0.073	0.049	0.122
		(0.037)		(0.069)	(0.073)	(0.079)
Reciprocal		0.003		-0.060*	-0.039	-0.088**
		(0.043)		(0.032)	(0.048)	(0.036)
Number of obs.	1,467	1,007	1,467	1,007	553	454
Number of subjects	156	110	156	110	110	110
Number of groups	12	8	12	8	8	8
Pseudo R-squared	0.267	0.264	0.246	0.220	0.189	0.287

Probit estimates, reporting marginal effects. Robust standard errors clustered at the group level in parentheses. Coefficient and s.e. of interaction term corrected, see f.n. 10. All specifications include the treatment order as a control variable. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Effort in Control Treatment

	(1)	(2)	(3)	(4)
	joint project		own project	
Bonus	0.846***	0.865***	-0.026	-0.048
	(0.029)	(0.042)	(0.071)	(0.090)
Good signal	0.558***	0.610***	0.732***	0.689***
	(0.076)	(0.108)	(0.082)	(0.104)
Bonus X good signal	-0.361***	-0.389***	-0.053	-0.034
	(0.077)	(0.106)	(0.063)	(0.077)
Wage5	-0.213***	-0.257***	0.062	-0.024
	(0.051)	(0.054)	(0.063)	(0.057)
Wage10	0.190***	0.108***	-0.079**	-0.088**
	(0.052)	(0.039)	(0.034)	(0.041)
Female		-0.008		0.006
		(0.105)		(0.070)
Altruist		-0.003		-0.029
		(0.085)		(0.072)
Trusting		-0.001		-0.029
		(0.042)		(0.074)
Fair		-0.080		0.072
		(0.065)		(0.081)
Reciprocal		0.061		0.042
		(0.050)		(0.079)
Number of observations	936	660	936	660
Number of participants	156	110	110	110
Number of groups	12	8	12	8
Pseudo R-squared	0.435	0.467	0.376	0.342

Probit estimates, reporting marginal effects. Robust s.e. clustered at the the group level in parentheses. Coefficient and s.e. of interaction term corrected, see f.n. 10. All specifications include the treatment order as a control variable. *** p<0.01, ** p<0.05, * p<0.1

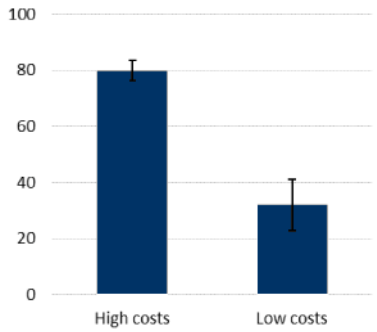

Figure 1: Mean bonus by observed level of costs (Main treatment). Error bars: +/- 2 SE.

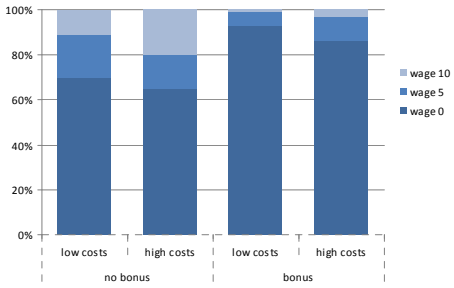

Figure 2: Distribution of the fixed wage by bonus.

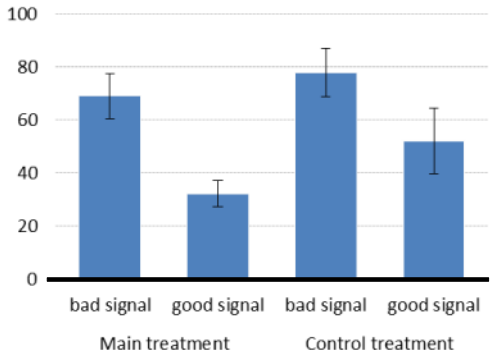

Figure 3: Difference in mean effort in the joint project between bonus and no bonus. Error bars: +/- 2 SE.

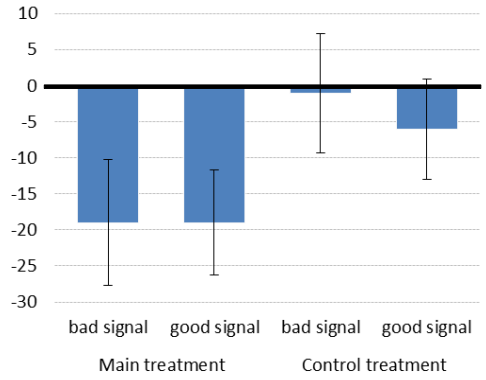

Figure 4: Difference in mean effort in the own project between bonus and no bonus. Error bars: +/- 2 SE.

Figure 5: Difference in mean effort in the own project between bonus and no bonus by round in the main treatment (3-round moving average).

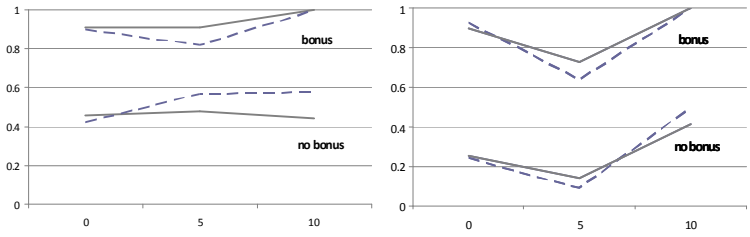

Figure 6: Mean effort in the joint project by wage level. Left panel: main treatment; right panel: control treatment. Solid lines is for all participants, dashed lines for reciprocal participants.