Walkowitz, Gari; Gürtler, Oliver; Wiesen, Daniel

Conference Paper

Behaving kindly, talking about it, and being rewarded for it?!

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Behavioral Economics, No. D10-V4

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Behaving kindly, talking about it, and being rewarded for it?!*

Oliver Gürtler†, Gari Walkowitz‡, Daniel Wiesen§

VERY PRELIMINARY: PLEASE DO NOT CITE OR CIRCULATE!!!

Abstract

In a principal-agent setup, we investigate agents' disclosure of conflict of interests—revealing deliberate or undeliberate kindness—and its affect on principals' reciprocal behavior. To this end, we firstly introduce a theoretical model refering to Hart and Moore (2008) which captures aspects of information revelation and reciprocal behavior. Secondly, a laboratory experiment ($N = 444$) tests behavioral predictions derived from the model. In the experiment, nature randomly determines the agent's choice set: either the agent can deliberately choose to behave kindly towards the principal (conflict of interest situation) or behaving kindly is the default. In any case, the agent can inform the principal about the available choice set. The principal can reciprocate the agent's behavior. We find agents to reveal their state when they are deliberately kind. Moreover, revealing a conflict of interest situation strongly triggers further reciprocal behavior by the principal. Our findings are robust towards different parameter variations. Implications are discussed.

JEL-Classification: C91, D82

Keywords: Asymmetric information, strategic information revelation, conflict of interest, reciprocity, laboratory experiment

*We are grateful for valuable comments and suggestions by Jeannette Brosig-Koch, Robert Dur, Bernd Irlenbusch, Patrick Kampkötter, Dirk Sliwka, Joel Sobel, Matthias Sutter, Joël van der Weele, and seminar participants at University of Cologne and the University of Oslo. We thank Emanuel Castillo and Katrin Recktenwald for their programming assistance, Rebecca Habisch for her excellent research assistance, as well as Anja Bodenschatz and Lisa Zander for their support in conducting experiments. Financial support of the German Science Foundation (DFG) through the Research Unit "Design & Behavior" is gratefully acknowledged.

†Faculty of Management, Economics and Social Sciences, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany; e-mail: oliver.guertler@uni-koeln.de.

‡Department of Corporate Development and Business Ethics, University of Cologne; e-mail: gari.walkowitz@uni-koeln.de.

§Department of Personnel Economics and Human Resource Management, University of Cologne; e-mail: wiesen@wiso.uni-koeln.de.

1 Introduction

Numerous field studies and lab experiments indicate that reciprocity is an important driver of human behavior (see, e.g., Fehr et al., 1993; Gneezy and List, 2006; Falk, 2007). In particular, people tend to be kind to other people that used to treat them well in the past. In many situations, however, people may find it difficult to assess whether they have been *really* treated well by someone else. Assessments might be particularly difficult in the absence of information on the available choice alternatives.

To illustrate this point, consider a person B that has received a payoff x as a consequence of an action by another person A. Suppose that B considers x to be rather high. Should B reciprocate by taking an action that makes A better off? Even though B feels that x is rather high, the answer to the above question may be negative. This is because B may not know whether A could have chosen an alternative action that would have benefited A, but which would have reduced B's payoff below x. In other words, because A's action set is unobservable to B, B does not know whether A really treated him kindly or whether A simply had to be kind by default.

The current paper investigates how an uninformed principal B reacts to information about the actions available for an informed agent A and whether A should in turn inform B about the action set. In case information provision is exogenous (i.e. beyond the control of A), one should expect that B feels treated well when learning that A took an action to increase B's payoff at the expense of his own payoff. It is then likely that B reciprocates by taking an action that benefits A. Things are not that clear when information is endogenously provided by A. On one hand, one may expect a similar argument to be true: whenever B learns that A treated him kindly, B may reciprocate by taking an action that makes A better off. However, an opposite reaction is also conceivable and the intuition is as follows. Suppose that A chose an action to increase B's payoff at his own expense. Assume further that A revealed his action set to B in order to demonstrate that he suffered to make B better off. While B may feel flattered by A's choice of action, the decision to reveal information about the action set indicates that A wants to be rewarded for his kindness. This makes the initial action appear less kind and may induce a spiteful reaction by B. In other words, the decision to reveal information about the available actions may be understood as an unkind act and may therefore backfire.

We develop a theoretical model to investigate the effects of endogenous information revelation on reciprocal behavior and tests the model predictions using data obtained from laboratory experiments. In the model, agent A can take an action that affects the own payoff and the payoff that another person B receives. The action set is randomly drawn by nature. With a certain probability, A has two available actions: one action that is particularly beneficial to himself and another one that is preferred by B. With the complementary probability, only

the latter action is available. After having taken an action, A can reveal the action set to B. Finally, B may pay a reward to A. We assume that B has intention-based preferences and is willing to reward behavior that he interprets as being kind. The strength of these preferences is captured by a parameter θ (preferences are modeled similar to those in Hart and Moore, 2008). We demonstrate that, depending on θ, one of two mutually exclusive equilibria is played. Whenever θ is low, A always takes the action that maximizes the own payoff and never reveals any information about the action set to B. Instead, when θ is high, A takes the action that maximizes B's payoff and reveals the action set whenever he could have taken the alternative action that would have increased the own payoff at B's expense. These results are intuitive. When θ is high, B is willing to pay a substantial reward to A when feeling kindly treated. As a consequence, A chooses the action that is best for B in order to induce a high reward. Moreover, to show that he deliberately increased B's payoff at the own expense, he reveals the action set when an alternative action that would have increased his payoff is available.

The model indicates that A can affect B's behavior in two ways. First, A can change the payoff accruing to B and therefore B's well-being. Ceteris paribus, when B receives a higher payoff, he feels treated more kindly and is therefore willing to reward A more generously. Second, A can inform B about the set of available actions. In this way, he can modify the payoff that B expects to receive. When B learns that A's action set includes alternatives that entail a low payoff for B, the payoff that he expects to receive is rather low. If his actual payoff is then rather high, he is very thankful to A and pays a higher reward in return.We employ laboratory experiments with direct-response and strategy-method treatments to test behavioral predictions of our model.

In our laboratory experiment, we use a neutral framing of the one-shot decision situation described in our theoretical model. Subjects are either allocated to the role of the agent or the principal. The experiment comprises three stages. In the first stage, the agent's choice set is randomly determined: two states M and N occur with equal chance. Possible actions and payoffs are common knowledge. In state M, the agent's choice set comprises two alternatives x and y. Alternative x yields a higher payoff to the principal, whereas alternative y yields a higher payoff to the agent (conflict of interest). In N, there is only alternative x available (no conflict of interest). The agent is said to behave (i) deliberately kindly if he chooses x and not y in M, (ii) deliberately unkindly if he chooses y in M, and (iii) undeliberately kindly if he chooses x in N. In the second stage, the agent can decide whether to inform the principal about the randomly drawn state. Information revelation is costly for the agent. In the third stage, the principal is informed about the payoff for herself and for the agent resulting from the agent's decision. The principal can reward the agent's kind behavior by a monetary amount.

To test the robustness of our theoretical model, we vary experimental parameters systemat-

ically, i.e., agents' costs of information revelation and the spread of principals' payoffs resulting from alternatives x and y. We also elicit agents' first order and principals' second order beliefs as well as individual characteristics.

Our behavioral results reveal that agents who deliberately behave kindly tend to inform principals about it, i.e., agents inform principals about the availability of a worse alternative for the principal if they deliberately choose the favorable alternative for the principal (in a conflict of interest situation). Agents who deliberately behave unkindly and agents who undeliberately behave kindly do not tend to inform principals. The main results for the principals are as follows: In line with theoretical predictions, we find that principals reward agents who behave (deliberately and undeliberately) kindly. Principals do not reward agents at all who do deliberately behave unkindly, however. The principals grant highest rewards to agents who behave deliberately kindly and inform principals about it. Rewards of informed principals for agents who behaved kindly undeliberately and rewards of uniformed principals (be it deliberately or undeliberately) are significantly lower.

Behavioral data support the main findings from the model. In particular we demonstrate that clever information provision by the agent can significantly increase the reward that he receives from the principal. We therefore conclude that an individual behaving kindly should openly talk about the situation if a conflict of interest was involved.

The paper proceeds as follows. In Section 2, we introduce a theoretical model to analyze kind behavior, information revelation, and reciprocity. Section 3 describes the experimental design and procedure. In Section 4, we present behavioral results and robustness checks. Section 5 concludes.

2 The model

2.1 Model description

Consider a situation with a principal B (she) and an agent A (he), both risk neutral. We have three different stages: In the first stage, the agent has to choose an action a_Σ. Σ denotes the set from which the action can be chosen. It depends on a move of nature and is either $\Sigma = M = \{x, y\}$ or $\Sigma = N = \{x\}$, where $\Sigma = M$ occurs with probability $q \in (0, 1)$ and $\Sigma = N$ with probability $1 - q$. If $a_\Sigma = x$, the gross payoffs to B and A are $u_B(x)$ and $u_A(x)$, respectively. Similarly, if $a_\Sigma = y$, gross payoffs are $u_B(y)$ and $u_a(y)$. We assume $u_B(x) > u_B(y) > 0$, $0 < u_A(x) < u_A(y)$. This means that there is a conflict of interest between the parties in that $a_\Sigma = x$ is preferred by B, while A prefers to choose $a_\Sigma = y$.

Of course, the agent knows the set of possible actions. Instead, the principal does not

receive this information. In the second stage, however, A can inform B about the relevant action set at a cost $k > 0$. We assume k to be small, i.e., $k > 0$ and $k \to 0$. This means that the agent always decides to inform the principal about the state of nature whenever this leads to an increase in his payoff. Let I_Σ denote an indicator variable that equals 1 if the agent has decided to inform the principal about the move of nature (in case the action set is Σ) and zero otherwise.

Related to Hart and Moore (2008), we assume B to have some reference level of gross payoff, denoted by $\tilde{u}_B$, she expects to receive. If B learns for sure that $\Sigma = M$ (or $\Sigma = N$), we impose $\tilde{u}_B = u_B(y)$ ($\tilde{u}_B = u_B(x)$). If B does not learn for sure which state has occurred, but believes to be in state $\Sigma = M$ with probability p, we have $\tilde{u}_B = pu_B(y) + (1-p)u_B(x)$. In words, we assume that the principal expects to receive the worst feasible outcome (or the expectation of the worst feasible outcome if she does not learn the state of nature for sure). Note that the principal uses all available information (e.g., the observation of the agent's choices in stage 1 and 2, but also the strategy she anticipates him to play in equilibrium) to determine p. If her actual payoff exceeds $\tilde{u}_B$, the principal feels happy. In this case, she can reciprocate by choosing some action $b \in \mathbb{R}_+$ in the third stage that increases the agent's payoff.[1] To account for happiness and reciprocation in the model, we follow Hart and Moore (2008) and write the two players' net utilities as[2]

$$\begin{aligned} U_A &= u_A(\cdot) + b - I_\Sigma \cdot k, \\ U_B &= u_B(\cdot) - |\theta(u_B(\cdot) - \tilde{u}_B) - b|. \end{aligned} \tag{1}$$

Consider the second term in U_B. If the principal gets a payoff above her expectation, she encounters a disutility of $\theta(u_B(\cdot) - \tilde{u}_B) > 0$ (e.g., because of a bad conscience) which she can reduce by increasing b.[3] The parameter $\theta > 0$ measures how strongly the principal reacts on a kind action by the agent.

We assume that neither the set of actions nor the actions themselves are verifiable so that the parties cannot write a contract to affect decisions.

[1] For example, the principal may increase the agent's payoff by recommending other principals to consult the agent.

[2] Note that Hart and Moore (2008) model expectations from an opposite perspective: they assume the principal to expect the best possible treatment. Moreover, they assume the principal to reduce the agent's payoff if her expectation is not met. Notice that our results would be very similar if we would follow this alternative modeling approach.

[3] Note that it is not possible to have $u_B(\cdot) - \tilde{u}_B < 0$. To see this, observe that the condition could only be fulfilled if the principal's gross payoff were $u_B(y)$. Then, however, the principal knew for sure that $\Sigma = M$ so that $\tilde{u}_B$ would be equal to $u_B(y)$ as well.

2.2 Model solution

The model is solved by backward induction. We begin with the principal's decision in stage 3. Here, it is straightforward to see that she chooses $b = \theta\left(u_B\left(\cdot\right) - \tilde{u}_B\right)$. As indicated before, she rewards the agent (i.e., $b > 0$) if her actual payoff exceeds her expectation ($u_B\left(\cdot\right) - \tilde{u}_B > 0$). Note that the agent can influence the principal's action in two ways. He can either change the principal's actual payoff $u_B\left(\cdot\right)$ through the action choice in stage 1 or her reference utility $\tilde{u}_B$ by informing her about the state of nature in stage 2. Let us analyze the agent's behavior in more detail. Although we have assumed his decisions to be sequential, from a game theoretic perspective we can treat them as being simultaneous. A strategy for the agent is a quadruple (a_M, I_M, a_N, I_N) specifying his action and revelation decision if $\Sigma = M$ and $\Sigma = N$. The agent can choose between $2 \cdot 2 \cdot 1 \cdot 2 = 8$ strategies. It is easy to show, however, that some of these strategies will never be played in equilibrium.

Lemma 1. The agent will never play a strategy involving $I_N = 1$.

Proof. Let $\Sigma = N$. Informing the principal about the state of nature could never decrease her reference utility and, as a result, would never lead to an increase in b. Since informing the principal also entails costs $k > 0$, this cannot be profitable for the agent. □

Lemma 1 is intuitive. If $\Sigma = N$, the agent is forced to choose $a_N = x$. Still, he wants to pretend that his action set were $\Sigma = M$ and that he voluntarily decided to choose the principal's preferred action. Lemma 1 simplifies the situation by eliminating four of the eight equilibrium candidates. Moreover, it indicates that there are two different kinds of (pure-strategy) equilibria. On the one hand, there may be an equilibrium where the agent informs the principal about the state of nature if $\Sigma = M$, i.e., $I_M = 1$. Such equilibrium can be understood as a *separating equilibrium* since the agent chooses a different action in stage 2 depending on the state of nature and, thus, the principal can infer from the agent's choice which state of nature is relevant. Similarly, there may be a *pooling equilibrium* where the agent chooses $I_M = 0$ so that his action in stage 2 is completely uninformative.

Let us analyze the two kinds of equilibrium one after the other. We begin with the *separating equilibrium*. Here, the agent chooses $I_M = 1$. If he also chooses $a_M = y$, his net payoff is $U_A = u_A\left(y\right) - k$. Similarly, if he chooses $a_M = x$, his net payoff is $U_A = u_A\left(x\right) + \theta\left(u_B\left(x\right) - u_B\left(y\right)\right) - k$. Hence, depending on whether or not $u_A\left(y\right) > u_A\left(x\right) + \theta\left(u_B\left(x\right) - u_B\left(y\right)\right)$, the choice of $a_M = y$ or $a_M = x$ is optimal. To see whether a separating equilibrium with the described features exist, we must show that the agent has no incentive to deviate from the specified strategy. If $u_A\left(y\right) > u_A\left(x\right) + \theta\left(u_B\left(x\right) - u_B\left(y\right)\right)$, we have just argued that it is optimal for A to choose $a_M = y$. Then, however, the principal infers the state of nature from the

observation of $a_M = y$ so that there is no need for the agent to reveal that information. Stated differently, if choosing $a_M = s$, the agent would always prefer to deviate to $I_M = 0$. If $u_A(y) < u_A(x) + \theta(u_B(x) - u_B(y))$, the agent chooses $a_M = x$. If he would deviate to $I_M = 0$, the principal would think that $\Sigma = N$ (since in the separating equilibrium only then information is withheld). Hence, she would think that the agent had no other choice than to play $a = x$ and would not reciprocate. Instead, by choosing $I_M = 1$, the agent can demonstrate that the principal's preferred action was chosen voluntarily which would be followed by B choosing $b = \theta(u_B(x) - u_B(y))$. This advantage of revealing information always outweighs the revelation costs k as these costs are assumed to be very small. The agent could also deviate from the separating equilibrium by changing the actions in both, stage 1 and stage 2 of the model. Then, he would choose $a_M = y$ and $I_M = 0$ instead of $a_M = x$ and $I_M = 1$. Accordingly, he would receive a net payoff $u_A(y)$ instead of $u_A(x) + \theta(u_B(x) - u_B(y)) - k$. Comparing these payoffs, the following proposition is immediate:

Proposition 1. If $\theta(u_B(x) - u_B(y)) - k \geq u_A(y) - u_A(x)$, there exists a separating equilibrium where the agent plays $(x, 1, x, 0)$, i.e., where the agent always chooses action x in the first stage, but informs the principal about the state of nature only in the case $\Sigma = M$.

The proposition indicates that existence of a separating equilibrium crucially depends on the strength of the principal's reward in case the latter feels well-treated (i.e., on θ and $u_B(x) - u_B(y)$). Only if the reward is sufficiently strong, the agent always chooses the principal's preferred action x. Then, to indicate that this choice was made voluntarily, he informs the principal about the action set whenever the alternative action is available.

Let us now turn to the *pooling equilibrium*, where $I_M = 0$. As seen before, if the agent chooses $a_M = x$, he has an incentive to choose $I_M = 1$ to indicate that he chose the principal's preferred action voluntarily. Hence, we can focus on the action $a_M = y$ when searching for a pooling equilibrium. If he chooses $a_M = y$, it is optimal for him to choose $I_M = 0$ so that the agent does not want to deviate with respect to the informational policy alone. If he thinks about deviating from $a_M = y$ to $a_M = r$, however, he always would want to choose $I_M = 1$ (as explained before). If the agent chooses $a_M = y$ and $I_M = 0$, his net payoff is again equal to $u_A(y)$. If deviating to $a_M = x$ and $I_M = 1$, his payoff would change to $u_A(x) + \theta(u_B(x) - u_B(y)) - k$. This means that the same payoffs as in the determination of the separating equilibrium are relevant. Thus, we obtain the following proposition.

Proposition 2. If $\theta(u_B(x) - u_B(y)) - k \leq u_A(y) - u_A(x)$, there exists a pooling equilibrium where the agent plays $(y, 0, x, 0)$, i.e., where the agent always chooses the action maximizing his own gross payoff and never informs the principal about the state of nature.

As argued before, the agent's equilibrium strategy and, hence, the resulting equilibrium depends on the strength of the principal's reward in case the latter feels well-treated. A pooling

equilibrium where the agent always chooses his preferred action (whenever it is available) and never informs the principal about the state of nature results if the reward is rather moderate.

Finally, it should be noted that reciprocity models other than that of Hart and Moore (2008) should lead to very similar conclusions. In the models by Rabin (1993), Levine (1998), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006), for example, the preferences of the players depend on their beliefs about motives or types of other players. Players want to be nice to players who treated them fairly or who are nice persons themselves. Applied to the decision problem we study this logic implies that A can affect the beliefs about his motives in a favorable way by revealing information to B about his action set when $\Sigma = M$. In turn, B should reward A particularly strongly if A chooses $(a_M, I_M) = (x, 1)$. This is easily illustrated using the model by Falk and Fischbacher (2006). On page 300, they define an *intention factor* which measures whether some player j had the opportunity to lower some other player i's payoff in case i gets a higher payoff than j. The intention factor is high if player i receives a higher payoff than j although j could have chosen some alternative action that would have entailed a lower payoff to i. Revealing information about the action set when $\Sigma = M$ and $a_M = x$ indicates to B that A could have chosen the alternative action y, which would have lowered B's payoff. As a consequence, the intention factor should increase which in turn would change B's preferences such that he would be more generous towards A.

The models mentioned above all study intention based reciprocity and interdependent preferences. There are also models of social preferences which assume that players care about their own payoffs and the payoffs to other people (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). According to these models, it does not matter how a specific pattern of payoffs is generated. Hence, revealing information about his action set would never be beneficial to A since it would not affect B's decision. Because this prediction contrasts starkly with the prediction from our model, our experiment can be used to discriminate between the different models of other-regarding preferences.

2.3 Behavioral hypotheses

We now pose a set of behavioral hypotheses for the agent's and principal's behavior according to the propositions of our theoretical model. These hypotheses are analyzed in our experiment.

First, we consider hypotheses for the agent's behavior. Assuming the agent behaves kindly (i.e., chooses x) in M, we state the following hypothesis about his information revelation to the principal:

Hypothesis 1. When A chooses x in state M, he informs B about the relevant state.

Further, we hypothesize for agents' information revelation to the principal:

Hypothesis 2. When A chooses y in state M or when the relevant state is N, A does not inform B about the relevant state.

Now, we focus on the impact of payoff-related parameter variations for the agent's behavior. According to our theoretical model, a decrease in the cost of information revelation k and an increase in the principal's gross payoff $u_B(x)$ makes it more likely that agent behaves kindly in M and informs the principal about the state. Thus, we state the following two hypotheses:

Hypothesis 3. The lower k, the more likely A is to choose x and to inform B about the state in state M.

Hypothesis 4. The higher $u_B(x)$, the more likely A is to choose x and to inform B about the state in state M.

Second, we state a set of hypotheses for the principal. According to our model, we hypothesize that the principal rewards the agent's kind behavior (i.e., choice of x) rather than unkind behavior (i.e., choice of y). This is in line with lab and field findings of reciprocal behavior in gift-exchange games, that are related to our setting (see, e.g., Fehr et al., 1993; Gneezy and List, 2006; Falk, 2007). Thus, we hypothesize:

Hypothesis 5. B rewards A for choosing x rather than y.

According to our model solutions on impact of the agent's information revelation on the principal's reward, we state the following hypothesis:

Hypothesis 6. The reward that A receives for choosing x is higher if B knows that the relevant state is M rather than if Bdoes not receive any information about the state or knows that the state is N.

Considering an increase in the principal's gross payoff, we hypothesize for the principal's reward of the agent's kind behavior:

Hypothesis 7. The reward that A receives for choosing x is increasing in $u_B(x)$.

3 Experimental design and procedure

3.1 General design and decision situation

In our experiment, we employ the direct-response method for a one-shot decision situation. A neutral framing is applied where each subject is either randomly allocated to the role of the agent (subject A) or the principal (subject B). One agent is randomly matched with one principal. In the direct-response experiment, we test the main behavioral hypotheses for the

agent (i.e., Hypothesis 1 and 2) and for the principal (i.e., Hypothesis 5 and 6).

Figure 1 illustrates the decision situation of the experiment. According to the theoretical model, the experiment comprises three stages. The structure of the game, possible actions and payoffs are common knowledge for the agent and the principal.

In the first stage, the agent's choice set is determined by a random draw. The two states

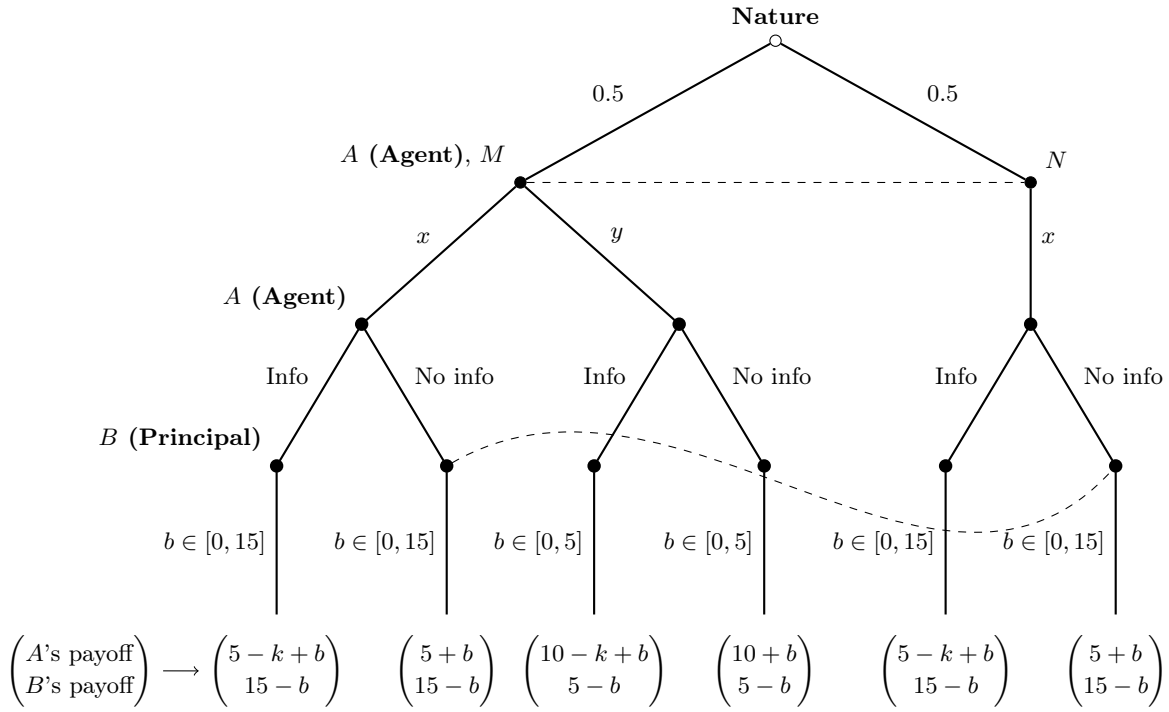

Figure 1: Illustration of the experimental decision situation

Notes: The figure illustrates the sequential decision situation of the experiment. Notice that if A chooses in state M, alternative x and alternative y principals' options to reward the agent are $b \in [0, 15]$ and $b \in [0, 5]$, respectively. The dotted line connecting states M and N indicates that knowing the state is A's private information. If A chooses x first stage and does not inform B the second stage, B cannot disentangle whether the state was M or N, a fact indicated by the bended dotted line.

M and N occur with equal probability ($q = 0.5$); whether the state is M or N is the agent's private information. In state M, the agent's choice set comprises alternatives x and y. The agent needs to choose one of those. Here, a conflict of interest for the agent is involved. Choosing alternative x is kind, as x yields a higher gross payoff to the principal (i.e., 5 for the agent and 15 for the principal).[4] The choice of alternative y is unkind, however, as it implies a higher gross payoff to the agent (i.e., 10) and a lower gross payoff to the principal (i.e., 5). In light of the conflict of interest situation in state M choosing x and y is *deliberately kind* and *deliberately unkind* behavior, respectively. In state N, the agent's choice set only comprises

[4]Payoffs and costs are shown in EUR. In the experiment, we employed Taler as our experimental currency converted at a rate of 1 Taler = 1 EUR.

alternative x and, thus, choose the kind alternative by default. The agent is said to behave *undeliberately kindly.*

In the second stage, the agent decides whether to inform the principal about the state that determined the available action set or not; it is a binary choice: (Info, No info). If the agent decides to inform the principal, he has to bear cost of $k = 1$ otherwise $k = 0$.

In the third stage, the principal learns about the gross payoffs resulting from the agent's decision. The principal also learns about the state (M or N) if the agent chose to inform the principal in stage 2. The principal then rewards agent's behavior by choosing a reward b. The range of the principal's reward is determined by agent choices, in particular, $b \in [0, 15]$ and $b \in [0, 5]$ if the agent chooses x and y, respectively.

The agent's and the principal's payoffs depend on the agents choices and the principal's reward. Formally, the agents' and the principal's payoff is $U_A = u_A(.) - k + b$ and $U_B = u_B(.) - b$, respectively.[5]

We also elicit the agent's first order and the principal's second order beliefs. At the end of the second stage, the agent is asked to state his belief on the magnitude of the principal's reward b. Analogously, the principal is asked at the end of the third stage about her belief about the agent's expectations about the principal's reward. After the principal stated her second order belief, final payoffs are revealed.

3.2 Control treatments

Besides the direct-response method, we employed the strategy method in additional experimental sessions to investigate whether the cost k or the principal's gross payoff $u_B(x)$ influence subjects' behavior. We carefully designed our strategy-method treatments closely to the treatments using the direct-response method to account for possible behavioral differences.[6] To address Hypothesis 3 and 4 for the agent and Hypothesis 7 for the principal, we vary the levels of cost and gross payoff systematically using a 2×2 factorial between-subjects design. Factor levels are $k_0 = 1$, $k_1 = 0.05$ and $u_B(x)_0 = 15$, $u_B(x)_1 = 20$. The treatment with the combination $k_0 = 1$ and $u_B(x)_0 = 15$ corresponds to our original treatment.

In the strategy-method treatments, the agent makes four decisions. In particular, the agent decides in state M on the action x or y and for each of the three nodes in the second stage (i.e., M, x, M, y, and N, x) illustrated in Figure 1 whether to reveal information to the principal. The principal makes five decisions on the reward b in each of the nodes in the third stage of the game.

[5] For a detailed description of subjects' tasks in the experiment see the instructions in the Appendix.

[6] Behavioral differences might occur for experiments employing direct-response and strategy methods; see Brandts and Charness (2011) for an excellent survey.

The order of decision tasks is randomized on subjects' screens. One decision is randomly chosen for a matched principal-agent-pair to be relevant for their payoff. Agents' first-order and principals' second-order beliefs are elicited analogously in control treatments.

3.3 Procedural details

Overall, 444 students from the University of Cologne participated in six experimental sessions of our computerized laboratory experiment. In particular, 316 and 128 subjects participated in the in the direct-response method and the strategy method treatments, respectively (see Table 1). Students were recruited by the online recruiting system ORSEE (Greiner, 2004). The experiment was programmed with zTree (Fischbacher, 2007).

The procedure of the experiment was as follows: upon arrival, subjects were randomly

Table 1: Experimental sessions

Session	Method	Treatments: factor levels	Number of subjects
1 to 10	Direct-response method	$k_0 = 1$, $u_B(x)_0 = 15$	316
11	Strategy method	$k_0 = 1$, $u_B(x)_0 = 15$	32
12	Strategy method	$k_1 = 0.05$, $u_B(x)_0 = 15$	32
13	Strategy method	$k_0 = 1$, $u_B(x)_1 = 20$	32
14	Strategy method	$k_1 = 0.05$, $u_B(x)_1 = 20$	32

Notes: This table indicates employed methods, factor levels, and number of subjects in the main experiment (i.e., sessions 1 to 10) and the control treatments (i.e., sessions 11 to 14).

allocated to the cubicles in the lab. Then, the experimenter provided some general information about the experimental procedure. Subjects was given plenty of time to read the instructions and for clarifying questions which were asked and answered in private. To check for subjects' understanding of the experiment we asked them to answer several comprehension questions. The experiment was not started unless all participants had answered the test questions correctly.

The decision task of the experiment lasted for about 15 minutes and 20 minutes in the direct-response method and the strategy-method treatments, respectively. In conjunction with the decision task, we elicited agents' first-order beliefs regarding principals' reward; likewise, we elicited principals' second-order beliefs, i.e., principals beliefs regarding the agents' expectation of rewards dependent on the agents' choices. Afterwards, subjects were asked to complete a comprehensive questionnaire comprising a psychometric measure to elicit subjects'

perception of their own integrity[7] and some demographic questions. Within the scope of our experiment, in addition to eliciting beliefs measuring subjects' integrity scores may help us to better understand their behavior.

Overall, the experiment lasted for about 45 minutes. Subjects earned on average 9.1 EUR from the experimental task alone. For completing the comprehensive questionnaire, each subject received an additional 2.5 EUR.

4 Results

In the following, we present our behavioral results. First, we analyze agents' and principals' behavior in the direct-response experiment using non-parametric statistics. Then, we investigate how subjects' beliefs relate to their behavior. Finally, we test for the robustness of our results employing regression analyses. Here, we also investigate the robustness of our model predictions by analyzing the impact of parameter variations using data from the strategy-method treatments.

4.1 Agents' behavior

For starters, we analyze agents' choices of alternatives and their information revelation in the first stage of the experiment (see Table 2). States M and N occurred almost with the same frequency: M and N in 49.4% and 50.6% of the cases, respectively.

Considering agents choices in state M indicates that 18 agents (22.5%) choose alternative x. On the contrary, 62 agents (77.5%) choose alternative y granting the higher payoff. In state N, all 78 agents are constrained to choose x.

We now investigate agents' revelation behavior. In state M, 10 agents (55.6%), who choose alternative x, inform principals about the state. No agent who chooses y in M informs the principal. In state N, 61 agents (78.2%) and 17 agents (21.8%) inform and do not inform principals, respectively.[8]

According to Hypotheses 1 and 2, agents inform principals if they have chosen x in state

[7] Our integrity measure captures individual differences in the inherent value of principled conduct, the steadfast commitment to principles despite temptations or costs, and the unwillingness to rationalize unprincipled behavior. It is based on subjects' self reports applying an established psychometric measure from social psychology (Schlenker, 2008). Higher scores on the scale reflect stronger claims of being committed to ethical principles.

[8] Notice that the latter behavior cannot be explained within the confines of our theoretical model. When being in state N, revealing information to the principal is neither a strategy in the separating, nor in the pooling equilibrium. This behavior might be due to the experimenter demand effect (see, e.g., Zizzo, 2010).

Table 2: Number of agents' choices and revelation of information

	State		
Agents' decisions	M	N	Total
Alternative x	18	78	96
Info	10	17	27
No info	8	61	69
Alternative y	62	–	62
Info	0	–	0
No info	62	–	62
Total	80	78	158

M and do not inform principals if they have chosen y in M or x in N. Our behavioral data support the hypotheses. First, we find that the majority of agents choosing x in M informs the principal. In contrast, no agent choosing y in M reveals information. Second, a minority of agents informs the principal in N. Applying test statistics shows that significantly more agents choosing alternative x inform principals about the state, when being in state M compared to state N (Pearson's χ^2=8.246, $p = .004$).[9] In sum, we state our first result:

Result 1. *Agents who deliberately behave kindly do tend to inform principals about it, i.e., agents inform principals about the availability of a worse alternative for the principal if they deliberately choose the favorable alternative for the principal. Agents who deliberately behave unkindly and agents who undeliberately behave kindly do not tend to inform principals about it.*

4.2 Principals' behavior

First we describe principals' behavior on the aggregate. Principals' reward agents with, on average, 1.27 (s.d. 2.31) in both states. This is 8.5% of the available amount.[10]

We now compare principals' rewards when agents choose alternative x and y. On average,

[9] All p-values reported throughout the paper are two-sided if not indicated otherwise.

[10] Recall that the available amount is either 15 or 5 dependent on whether agents choose x or y. As principals always choose $b = 0$ whenever agents choose y, the share of 8.5% only refers to the amount of 15 (see below). This share is substantially lower than giving rates in dictator games considering a share of 28.4% reported in Engel's (2011) meta study. If we consider only the surplus amount that the principal can distribute after pay-off equalization, i.e., pocketing 5 Talers and distribute only 10 Talers, the resulting share (12.7%) is still below the above reported threshold. In traditional trust games, responders typically back-transfer around 45% of the initial senders' investment (see Camerer, 2003).

96 principals reward agents who choose alternative x with 2.09 (s.d. 2.66). On the contrary, all 62 principals agents do not reward agents who choose alternative y. Obviously, principals' rewards differ significantly across alternatives ($p = 0.000$, Fisher-Pitman permutation test for independent samples).[11] This behavioral pattern is in line with Hypothesis 5 and, thus, we state the following:

Result 2. *Principals reward agents who behave (deliberately or undeliberately) kindly. Principals do not reward agents who do deliberately behave unkindly.*

We now investigate whether rewards of informed principals in M, informed principals in N, and uninformed principals differ for agents who choose x. Figure 2 shows that rewards of informed principals in M are substantially higher compared to rewards of uninformed principals (in M and N) and informed principals in N.

According to Hypothesis 6, principals' rewards for agents' choice of alternative x differ

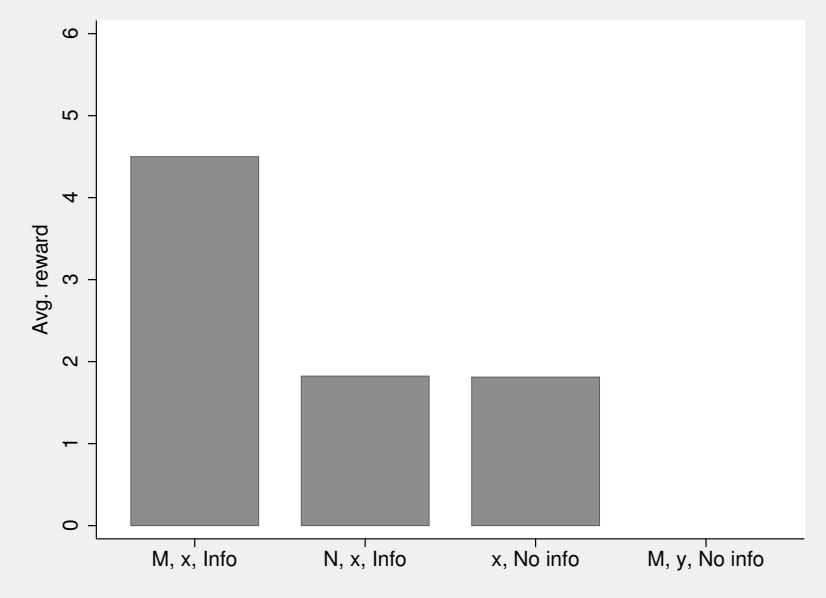

Figure 2: Heterogeneity in principals' rewards

Notes: This figure illustrates principals average rewards for the four occurring situations M, x, Info (N = 10), N, x, Info (N = 17), x, No info (N = 69), and M, y, No info (N = 62).

across states and with agents' information revelation. Intuitively, principals' rewards are supposed to be affected if agents reveal or not reveal that their kind behavior has been deliberate or not. Our results are in line with the above hypothesis. Informed principals in state M grant, on average, significantly higher rewards ($\bar{b} = 4.50$[12], s.d. 2.55) compared to informed

[11] In the following, Fisher-Pitman permutation tests for independent samples are denoted as FPTI.

[12] Note that this number exactly corresponds to the fraction typically reported for subjects' back-transfers

principals in N ($\bar{b} = 1.82$ s.d. 2.07; $p = 0.011$, FPTI). Informed principals in M also grant significantly higher rewards than uninformed principals ($\bar{b} = 1.81$, s.d. 2.65; $p = 0.012$, FPTI). The difference in rewards across informed principals in N and uninformed principals is not significant, however ($p = 0.973$, FPTI). In short, we can state the following result:

Result 3. *Principals grant highest rewards to agents who behave deliberately kindly and inform principals about it. Rewards of informed principals for agents who behaved kindly undeliberately and rewards of uniformed principals (be it deliberately or undeliberately) are significantly lower.*

Result 3 indicates that revealing information about deliberate kind behavior triggers further reciprocal behavior. Interestingly, principals reward the mere fact that kind behavior has been deliberate and not by default although payoffs would be equivalent. Whether agents used the option to reveal information "strategically" anticipating a higher reward in M, x when informing the principal, did not seem to matter for principals' rewards. Our result is also line with the experimental literature emphasizing the importance of intentions in fair behavior (see, e.g., Falk et al., 2003, 2008).

4.3 Beliefs

We now describe subjects' beliefs and investigate how beliefs could help us to explain subjects' behavior. In particular, we investigate agents' first-order beliefs about principals' rewards and principals' second order-beliefs (belief on the agents' expected reward). Table 3 provides descriptive statistics on agents' first-order beliefs and principals' second-order beliefs.

Overall, principals believe that agents expect, on average, a reward of 2.26. Informed principals belief that agents who behave kindly deliberately expect highest rewards. Further, informed principals belief that agents who behaved kindly undeliberately expect a reward of 3.06 which is substantially lower. Uniformed principals belief that agents who behave kindly expect a slightly higher reward of 3.88. Here, the principal is not able to distinguish whether an agent's kind behavior has been deliberate or not, which might explain the slightly higher second-order belief. Uninformed principals belief that agents who behaved unkindly do expect the lowest reward, being slightly larger than zero.

Overall, agents' expect an average reward of 1.79 from principals. Agents who behaved kindly and informed the principal about it belief that principals' reward will be highest (i.e., 5.00). Agents who behaved kindly but undeliberately and informed the principal about it, expect a substantial lower reward of 3.00. Agents who behaved kindly be it deliberately or

in trust games.

Table 3: Principals' second order and agents' first order beliefs

	Principals' beliefs			Agents' beliefs		
Situation	Mean	s.d.	N	Mean	s.d.	N
M, x, Info	4.80	1.40	10	5.00	2.00	10
N, x, Info	3.06	3.07	17	3.00	2.75	17
M, N, x, No info	3.88	2.32	69	2.46	2.56	69
M, y, No info	0.52	1.40	62	0.18	0.74	62
Total	2.26	2.77	158	1.79	2.47	158

Notes: This table shows descriptive statistics on principals' beliefs on agents' expected rewards for four different situations determined by states M and N and agents' behavior.

undeliberately without informing the principal expect a slightly lower reward, on average (i.e., 2.56). Among them are, in N, eight agents who behave kindly deliberately and expect a substantial higher average reward (i.e., 4.86, s.d. 0.35). Their belief indicates that their behavior could have been driven by, for example, false consensus or curse of knowledge considerations implying that information revelation becomes unnecessary. In M, 61 agents who behave kindly undeliberately expect an average reward of 2.15 (s.d. 2.56). As in N agents behave kind by default, the motive for their decision not to reveal information might be due to concealing that their kind behavior has been undeliberate. The considerably low expectation of principals' reward might also be driven by those individuals who avoid the cost of revealing information expecting non-reciprocal behavior by the principal. Agents who behaved deliberately unkind expect almost no rewards from uninformed principals.

Now we analyze how principals' beliefs are correlated with rewards. Overall, we find a significant correlation between principals' second order beliefs and assigned rewards (Spearman's $\rho = 0.490$, $p = 0.000$). Looking at correlations dependent on agents' state, chosen alternative and information revelation decision, we find a very strong correlation between principals' belief and reward in M, x, Info (Spearman's $\rho = 0.921$, $p = 0.000$). A smaller, yet still significant, correlation is observed when the agent chooses x but does not inform the principal about her state, i.e., in x, No info (Spearman's $\rho = 0.281$, $p = 0.019$). No statistically significant correlation can be found in N, x, Info.[13]

[13]Note that we cannot calculate feasible correlations for M, y, No info (due to the lack of variance in principals' rewards) and M, y, Info (due to the lack of empirical evidence).

4.4 Robustness of results

In the following, we run a series of regressions to analyze the robustness of our findings by considering control variables like subjects' beliefs and demographics. First, we will analyze agents' and principals' choices in our direct response treatment. Second, we will assess the influence of the cost k for information revelation and the principal's gross payoff $u_B(x)$ on agents' and principals' behavior. As predicted by our hypotheses, lowering k will strengthen the agent's inclination to choose x and to inform the principal about the state when her state is M. Moreover, an increase in $u_B(x)$ should lead to an increase in agents choosing x and inform the principal when they are in M. Finally, the principal's reward is predicted to increase in $u_B(x)$.

To start with, in Table 4 we regress agents' information revelation decision in N, x and M, x relative to M, y. Model (1) conveys that the probability that agents reveal their state is significantly higher in N, x and M, x, i.e., when agents behave (deliberately or undeliberately) kindly. Moreover, comparing the probability of revealing information in N, x and M, x, a Wald test shows that in M, x—when agents behave deliberately kindly—agents are significantly more likely (99.6%) to reveal their state compared to N, x (46.9%, $p = 0.007$)—when they behave undeliberately kindly.

This finding is robust when controlling for agents' belief about principals' rewards (model (2)), agents' sex and age (model (3)) and a measure for agents' integrity (model (4)). In sum, models (2) to (4) also show that none of the added control variables significantly contributes to the prediction of agents' information revelation choice.

We now investigate principals' rewards to agents more closely for the agents' state and information revelation decision. Table 5 depicts coefficients from OLS-regression analyses predicting principals' reward in N, x, Info and M, x, Info—when she was informed by the agent—and x, No info, when the principal faces x without information on the state x was chosen from.[14] Model (1) shows principals' rewards in N, x, Info, M, x, Info, and x, No info relative to M, y, No info, the reference category. Since M, y, No info, is always rewarded with $b = 0$, the coefficients from model (1) map the average transfers reported above. In all three cases (N, x, Info, M, x, Info, x, No info) principals assign the agent a significantly higher reward compared to M, y, No info. A series of Wald-tests further shows that principals reward agents' revealed deliberate kindness in M, x, Info significantly more compared to N, x, Info and x, No info ($p \leq 0.004$). Interestingly, we find no evidence that principals assign rewards differently across the latter two cases ($p = 0.984$).

Entering principals' second order belief, i.e., her belief regarding the agents' expectation on

[14] N, y, Info can be disregard from our analyses as the agent never informed the principal when choosing M, y.

Table 4: Probit analysis of agents' information revelation

	Direct-response experiment				Strategy-method experiment					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent variable:	Info	Info	Info	Info	Info	Info	Info	Info	Info	Info
M, x	0.996***	0.995***	0.994***	0.995***	0.607***	0.648***	0.555***	0.562***	0.551***	0.572***
	(0.003)	(0.005)	(0.003)	(0.013)	(0.0751)	(0.0748)	(0.0996)	(0.124)	(0.126)	(0.130)
N, x	0.469***	0.450***	0.434***	0.434***	0.402***	0.438***	0.338***	0.349***	0.337***	0.335***
	(0.058)	(0.071)	(0.066)	(0.096)	(0.085)	(0.088)	(0.110)	(0.108)	(0.108)	(0.111)
Factor k						-0.221**	-0.215**	0.253***	0.280***	0.329***
						(0.088)	(0.091)	(0.097)	(0.098)	(0.089)
Factor $u_B(x)$						0.167*	0.166*	0.204**	0.217**	0.284***
						(0.091)	(0.093)	(0.103)	(0.106)	(0.102)
First-order belief		0.000	0.000	0.000			0.030	0.030	0.034	0.038*
		(0.000)	(0.000)	(0.001)			(0.022)	(0.021)	(0.023)	(0.020)
$M, x\times$ Factor k								0.103	0.109	0.149
								(0.115)	(0.113)	(0.126)
$M, x\times$ Factor $u_B(x)$								-0.090	-0.090	-0.117
								(0.100)	(0.099)	(0.100)
Sex			-0.002	-0.002					-0.122	-0.198*
			(0.002)	(0.003)					(0.120)	(0.110)
Age			0.000	0.000					-0.002	-0.008
			(0.000)	(0.000)					(0.008)	(0.009)
Integrity				-0.000						0.399***
				(0.001)						(0.108)
Observations	158	158	158	158	192	192	192	192	192	192
Pseudo-R^2	0.263	0.273	0.284	0.284	0.185	0.242	0.268	0.271	0.280	0.336

Notes: The dependent variable in all columns is the agents' information revelation. Information revelation is coded 1; no information revelation is coded 0. The estimation method in all columns is probit regressions. Coefficients are shown as marginal effects. The reference category is M, y. Robust standard errors are shown in parentheses under the coefficients. Wald test results indicating a significant difference between M, x and N, x ($p \leq 0.0389$).

***Significant at a 1 percent level.

**Significant at a 5 percent level.

*Significant at a 10 percent level.

principals' reward dependent on the agent's state and information revelation decision, indicates that second-order beliefs predict principals' rewards. With other words, principals appear to be sensitive toward their belief on the agents' expectation (model (2)). In model (3), we add an interaction term of M, x, Info, and principals' second-order belief. The model shows that the reward-enhancing effect of principals' second order belief depends on the state the agents chooses x from and on whether she informs the principal about the state. In models (4) and (5), we control for principals' sex, age and integrity. Both models confirm the findings from models (1) to (3). In addition, females and subjects scoring high in integrity assign higher rewards to agents.

Our behavioral results for agents are robust towards variations in the experimental method and experimental parameters. When we compare models (1) to (4), using data from our direct-response experiments, with models (5) to (10) from the strategy-method treatments, in Table 4, we find quite similar estimation results for agents' information revelation decision across treatments. To test hypotheses 3 and 4, we assess whether an increased cost for information revelation and an increase in principal's gross payoff affect agents' information revelation decision negatively or positively, respectively, by controlling for different levels of k and $u_B(x)$ in models (6) to (10). We find evidence that agents' information revelation decision is negatively affected by increasing k and positively by increasing $u_B(x)$. In addition, a Wald-test shows that agents seem to be more sensitive towards the cost they have to bear when they reveal information as compared to an induced gross-payoff inflation for the principal ($p = 0.04$). Yet, we find no evidence for an interactive effect of lower k and higher $u_B(x)$, respectively, and M, x, as predicted by our hypotheses 3 and 4. Therefore, we cannot confirm that agents who are more likely to choose alternative x from state M–induced by lower k or higher $u_B(x)$–are also more likely to inform the principal about their state. Controlling for agents' first-order belief further conveys that agents do not condition their information revelation on their expected reward by the principal. Finally, model (10) shows that agents scoring high on integrity are more likely to reveal their state.

For principals, main behavioral results are also robust towards variations in the experimental method and experimental parameters. Comparing model (1) using data from our experiments with direct-response method with model (6) from the strategy-method treatment in Table 5 indicates very similar estimation results. According to Hypothesis 7, we test whether variations in principals' gross payoff $u_B(x)$ alter principals' rewards. To this end, we control for different levels of $u_B(x)$, employed in our factorial design, in model (7). Estimation results indicate no significant influence on rewards increasing $u_B(x)$. Also, a reduction in cost of information revelation k for the agent does not significantly affect principals' behavior.

Analogous to findings from the direct-response experiments, principals' second-order beliefs significantly affect principals' behavior; see model (8). This again suggests, that principals

Table 5: OLS-regression analyses of principals' rewards

	Direct-response experiment					Strategy-method experiment					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Dep. variable	Reward	Reward	Reward	Reward	Reward	Reward	Reward	Reward	Reward	Reward	Reward
M, x, Info	4.500***	3.742***	-3.880***	-4.528***	-3.689***	4.016***	4.016***	0.868	-0.092	-0.132	0.144
	(0.775)	(0.849)	(0.921)	(1.044)	(0.988)	(0.309)	(0.310)	(0.603)	(0.711)	(0.724)	(0.695)
N, x, Info	1.824***	1.228**	1.338**	1.261**	1.183**	2.906***	2.906***	0.262	0.426	0.434	0.605
	(0.493)	(0.556)	(0.546)	(0.555)	(0.522)	(0.304)	(0.305)	(0.539)	(0.585)	(0.585)	(0.595)
x, No Info	1.812***	1.362***	1.445***	1.389***	1.317***	2.484***	2.484***	0.420	0.548	0.555	0.688
	(0.321)	(0.292)	(0.289)	(0.286)	(0.271)	(0.303)	(0.304)	(0.457)	(0.497)	(0.501)	(0.515)
M, y, Info						0.047*	0.047*	0.047	0.047	0.047	0.047
						(0.027)	(0.027)	(0.069)	(0.066)	(0.066)	(0.062)
SO-Belief		0.177*	0.144	0.132	0.162*			0.546***	0.512***	0.510***	0.475***
		(0.096)	(0.090)	(0.090)	(0.089)			(0.086)	(0.097)	(0.098)	(0.102)
SO-Belief × M, x, Info			1.617***	1.760***	1.614***				0.188	0.196	0.184
			(0.191)	(0.207)	(0.199)				(0.141)	(0.144)	(0.143)
Factor k							-0.086	-0.094	-0.099	-0.132	-0.106
							(0.359)	(0.286)	(0.286)	(0.291)	(0.270)
Factor $u_B(x)$							0.475	-0.323	-0.368	-0.393	-0.310
							(0.359)	(0.289)	(0.289)	(0.287)	(0.284)
Sex				0.540*	0.474*					0.306	0.303
				(0.276)	(0.265)					(0.291)	(0.272)
Age				-0.010	-0.009					-0.009	-0.005
				(0.021)	(0.023)					(0.045)	(0.041)
Integrity					1.164**						1.208***
					(0.573)						(0.319)
Constant	0.000	-0.091	-0.074	-0.078	-3.986*	0.172	-0.022	0.167	0.205	0.281	-3.853**
	(0.000)	(0.056)	(0.051)	(0.556)	(2.065)	(0.111)	(0.276)	(0.226)	(0.219)	(1.089)	(1.493)
Observations	158	158	158	158	158	320	320	320	320	320	320
Adjusted R^2	0.261	0.289	0.341	0.347	0.383	0.400	0.406	0.559	0.561	0.562	0.588

Notes: The table displays coefficients from ordinary least square-regression models. The reference category is M, y, No Info. Robust standard errors are shown in parentheses under the coefficients.

***Significant at a 1 percent level.

**Significant at a 5 percent level.

*Significant at a 10 percent level.

decision to reward depends on their belief regarding the agent's expectation concerning principals' rewarding behavior. Adding principals' second-order beliefs to the regression also implies that main effects are not significant anymore. Moreover, contrary to our direct response experiment (models (3) to (5)), we do not find an interactive effect between principals' second order-belief and M, x, Info (models (9) to (11)). Including further controls for demographics does not have a significant effect (model 10). Similar to model (5), model (11) indicates that principals' stated integrity measure positively affects principals' rewarding behavior.

5 Concluding remarks

The present paper analyzes the revelation of kind behavior in a conflict of interest situation and its impact on reciprocal behavior. To this end, we firstly develop a theoretical principal-agent model to investigate the effects of endogenous information revelation on reciprocal behavior and secondly test behavioral predictions using data from laboratory experiments.

In line with our theoretical predictions, our behavioral results reveal that agents who deliberately behave kindly tend to inform principals about it, i.e., agents inform principals about the availability of an inferior alternative for the principal if they deliberately choose the favorable alternative for the principal (in a conflict of interest situation). Contrarily, agents who deliberately behave unkindly and agents who undeliberately behave kindly do not tend to inform principals. The main results for the principals are as follows: In line with theoretical predictions, we find that principals reward agents who behave (deliberately and undeliberately) kindly. Principals do not reward agents at all who do deliberately behave unkindly, however. These findings are similar to lab and field evidence on reciprocal behavior (see, e.g., Fehr et al., 1993; Falk, 2007). The principals grant highest reward when they know that the agent behaved deliberately kindly, i.e., when the agent informed the principal about it. Rewards of informed principals for agents who behaved kindly undeliberately and rewards of uniformed principals (be it deliberately or undeliberately) are significantly lower. In terms of payoffs, agents who deliberately behaved kindly and informed the principal about it (at a cost) are, on average, even over-compensated for their behavior.

In sum, our results show that revealing deliberate kindness triggers strong reciprocal behavior. Even at costs and in a one-shot interaction, information revelation seems beneficial. Intuitively, we conclude that an individual behaving kindly should openly talk about the situation if a conflict of interest was involved and when principals face asymmetric information about agents' action set.

References

Bolton, G. E. and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Brandts, J. and G. Charness (2011): "The strategy versus the direct-response method: a first survey of experimental comparisons," *Experimental Economics*, 14, 375–398.

Camerer, C. (2003): *Behavioral Game Theory*, Princeton (NJ), Princeton University Press.

Dufwenberg, M. and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268 – 298.

Engel, C. (2011): "Dictator games: A meta study," *Experimental Economics*, 14, 583–610.

Falk, A. (2007): "Gift Exchange in the Field," *Econometrica*, 75, 1501–1511.

Falk, A., E. Fehr, and U. Fischbacher (2003): "On the Nature of Fair Behavior," *Economic Inquiry*, 41, 20–26.

——— (2008): "Testing theories of fairness—Intentions matter," *Games and Economic Behavior*, 62, 287 – 303.

Falk, A. and U. Fischbacher (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293 – 315.

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

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

Fischbacher, U. (2007): "Z-tree: Zurich Toolboox for Readymade Economic Experiments – Experimenter's Manual," *Experimental Economics*, 10, 171–178.

Gneezy, U. and J. A. List (2006): "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, 74, 1365–1384.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen : Beiträge zum Heinz-Billing-Preis 2003*, ed. by K. Kremer and V. Macho, Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen.

Hart, O. and J. Moore (2008): "Contracts as Reference Points," *Quarterly Journal of Economics*, 123, 1–48.

Levine, D. K. (1998): "Modeling Altruism and Spitefulness in Experiment," *Review of Economic Dynamics*, 1, 593–622.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83, 1281–1302.

Schlenker, B. R. (2008): "Integrity and character: Implications of principled expedient ethical ideologies," *Journal of Social and Clinical Psychology*, 27, 1078–1125.

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

Appendix

A Instructions

[Adaptations for strategy-method treatments displayed in brackets]

General information

You are now taking part in an economic decision experiment. Please read the following instructions for the experiment carefully. Throughout the entire experiment it is very important that you do not talk to any of the other participants. In case you do not understand something, please read the corresponding instructions again. If you then still have questions, please raise your hand. We will come to your cabin and answer your question personally.

In this experiment you can earn money. The earnings you receive during the experiment are specified in Taler. The amount of Taler you can earn during the experiment depends on your decisions and on the decisions of one other participant.

All Taler earned will be converted into Euros at the end of the experiment. For this it holds: 1 Taler = 1 EURO.

During the experiment, you interact with one randomly assigned other participant. You are at no point in time informed about the name of the other participant. Likewise the other participant does not get to know your identity at any point in time.

All data and answers will be analyzed anonymously. In order to assure anonymity, you have drawn a personal code. We can only match your decisions to this code but not to you as a person.

After the experiment, we will ask you to fill in a questionnaire which we need in addition for a statistical analysis.

Experimental procedure

In the experiment, two types of persons make their decisions: **Person A** and **Person B**. In the beginning, it will be randomly drawn, if you will decide as a Person A or as a Person B. Also, there will be a random matching of one person A with one person B. The matched persons interact in the experiment. Both, your role (A or B) and the matching with the other person, will be the same throughout the experiment.

The amount in Taler which Person A earns during the experiment depends on his decisions and the decision of Person B. In the same way, the amount of Taler for Person B will be determined by his decisions and the ones of Person A.

[In the following, the general structure of the decision situation of Person A and Person B is described.]

The experiment consists of two stages, which will be described in the following:

Stage 1: Decisions of Person A
In the beginning of stage 1, the computer randomly chooses one of two states. This determines Person A's set of choice alternatives. The chance for both states to be drawn is the same (50%): **State M or state N**. Thereafter, Person A selects one of the alternatives in each state. This choice influences his own payoff and the payoff of the matched Person B.

- **State M:** In state M, Person A has two choice alternatives: **Alternative x** and **Alternative y**. In the case Person A chooses Alternative x, Person A receives 5 Taler and Person B receives 15 Taler. If Person A chooses Alternative y, he receives 10 Taler and Person B receives 5 Taler.
- **Sate N:** In state N, there is only one choice alternative: **Alternative x**. Thus, for Person A it is only possible to choose alternative x. Person A now receives 5 Taler and Person B receives 15 Taler.

Please note: At this stage, the randomly drawn state (M or N) is only known to Person A. That means, only Person A knows the choice alternatives actually available.

The following table provides an overview about the states $M$and N, the choice alternatives x and y, and the payoffs for Person A and Person B:

	State M **with 50% chance**		State N **with 50% chance**
	Alternative x	**Alternative y**	**Alternative x**
Payoff of Person A	5 Taler	10 Taler	5 Taler
Payoff of Person B	15 Taler	5 Taler	15 Taler

Person A can inform Person B about the occuring state (M or N) after the choice of an

alternative. In case Person A decides to inform Person B, there will be cost of 1 Taler for Person A. If Person A decides not to inform Person B, there will be no cost. Afterwards, stage 1 is completed.

Stage 2: Decision of Person B
In the second stage, **Person B gets to know his or her payoff (15 or 5 Taler)** depending on Person A's choice in the first stage. Only if Person A decided to inform Person B about the state in the first stage, Person B learns in which of the two states (M or N) Person A had to decide.

Person B can make a payment to Person A. Person B has the amount at his disposal which he got from Person A after the first stage. Accordingly, Person B can either choose an integer from **0 to 15 Taler (if Person A chose Alternative x)** or from **0 to 5 Taler (if Person A chose Alternative y)**. Afterwards, stage 2 is completed.

How to calculate total payoffs?

[In the experiment, at first you decide for all possible combinations of states (M and N), therewith associated alternatives (x and y) and (as Person B) for situations where Person A has either informed Person B or not. Which decisions you take depends on your randomly assigned type (either Person A or Person B). For payoff calculation, for every matched Person A and Person B dyad, a state is randomly determined. Afterwards, the decisions of Person A and Person B for that state are compared. Dependend on which alternative Person A has chosen and whether Person A has informed Person B in which of the two states M or N Person A was, the respective payment of Person B to Person A is realized.]

The payoffs for Person A and Person B are as follows:

Person A
Person A's total payoff depends, for one thing, on the choice of alternatives in stage 1 ($x = 5$ Taler or $y = 10$ Taler in M; $x = 5$ Taler in N). Also, the total payoff depends on the decision to inform Person B about the realized state or not (-1 Taler if Person A informs Person B about the realized state; 0 Taler if Person A does not inform Person B about the realized state). Finally, Person A's total payoff depends on the payment he gets from Person B in stage 2 ($0, 1, 2, 3, 4, 5$ Taler or $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$ Taler).

Person A's total payoff = payoff at stage 1 − costs in case of the decision to inform Person B + payment of Person B at stage 2.

Person B

The total payoff of Person B depends, for one thing, on the choice of alternatives in stage 1 by Person A ($x = 15$Taler or $y = 5$ Taler in M; $x = 15$ Taler in N). Also, the total payoff depends on the payment to Person A in stage 2 ($0, 1, 2, 3, 4, 5$ Taler or $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11$, 12, 13, $14, 15$ Taler).

Person B's payoff = payoff in stage 1 − payment to Person A in stage 2.