On the Hidden Costs of Incentive Schemes

Dirk Sliwka[*]
University of Cologne and IZA Bonn

May, 2003

Abstract

By enriching a principal-agent model it is shown that the introduction of monetary incentives may reduce an agent's motivation. In a first step, we allow for the possibility that some agents stick to unverifiable agreements. The larger the fraction of reliable agents, the lower powered will then be the optimal incentive scheme and fixed wages become optimal when performance measurement is costly. If social norms matter such that some agents' reliability is influenced by their beliefs on the convictions of others, high powered incentives signal that not sticking to agreements is a widespread behavior and may lead to lower effort levels.

Key Words: Incentives, Intrinsic Motivation, Motivation Crowding-Out, Honesty

JEL Classification: M52, J33, D23

[*] Seminar für allgemeine Betriebswirtschaftslehre und Personalwirtschaftslehre, University of Cologne, Herbert-Lewin-Str. 2, 50931 Köln, Germany, tel: +49 221 470-5888, fax: +49 221 470-5078, e-mail: dirk.sliwka@uni-koeln.de.

1 Introduction

Most economists claim that incentive contracts in firms are beneficial as they raise efficiency by increasing employees' motivation to work harder. Indeed there are some recent empirical studies showing for single firms that incentive contracts have raised productivity significantly.[1]

However, when looking on the frequency of occurrence of contracts based on individual performance evidence is less overwhelming in favor of incentives. Parent (2001) for instance surveys different samples of the US working population and from his summary statistics at most one quarter of all employees receives some form of compensation based on individual performance.[2]

Indeed there seem to be very different views in individual firms on whether contracts based on individual performance are beneficial or not. Some see incentive contracts as an important component of their human resource management practices whereas others take a much more sceptical view and may even consider individual extrinsic incentives as harmful.

Psychologists have for quite some time also taken a much more sceptical view on extrinsic incentives than most economists. Since the seminal work by Deci (1971), it has often been pointed out that monetary incentives can be harmful as they may reduce *intrinsic motivation.* Very roughly two strands of arguments are given for this *motivation crowding-out effect.* On the one hand, there is cognitive evaluation theory, positing that individuals strive for competence and self-determination and monetary rewards undermine self-determination and therefore possibly the joy of performing the task. A different explanation is given on the basis of self-perception theories stating that individuals sometimes do not understand their own motives perfectly and – in the language of economics – learn boundedly rational from their own actions on their motives (or preferences). If monetary incentives are set for an activity, then an individual concludes that it performs this activity because of those incentives. If the incentives are abandoned, motivation is reduced as compared to a situation where there never have

[1] Compare for instance Lazear (2000) and the overviews given by Gibbons (1997) or Prendergast (1999).

[2] This number encompasses piece rates, commissions and bonusses. When excluding bonusses the fraction drops to less than 10%. Compare also the summary statistics given in Parent and MacLeod (1999).

been extrinsic incentives.[3]

Numerous experimental studies have been conducted by psychologists on this issue. On balance, the evidence is mixed whether extrinsic incentives are beneficial or whether they may reduce motivation.[4] There are some recent economic experiments trying to capture the situation described by principal-agent models more closely. In laboratory experiments, Fehr and Gächter (2002) or Irlenbusch and Sliwka (2003) have observed that the introduction of a possibility to set incentives[5] made principals worse off and reduced efficiency as compared to a pure fixed wage setting. Interesting experiments with real efforts have been conducted by Gneezy and Rustichini (2000a), who find that setting weak monetary incentives may actually worsen results as compared to a pure fixed compensation for tasks such as answering questions taken from an IQ test or collecting for a charity. In a different paper, Gneezy and Rustichini (2000b) examined the introduction of a fine for parents when being late at picking up their children from day-care centers. Surprisingly, the introduction led to a significant increase in the number of late-coming parents. Field evidence in line with motivation crowding-out is found by Frey and Götte (1999), where a negative relationship between the time spent for volunteer work and the fact that there is monetary compensation for it is found with data from the Swiss Labor Force Survey.[6]

But how can this variety of standpoints and results observed be reconciled with the economics of incentives? Recently this issue has been taken up by some economists. Kreps (1997) for instance gives an informal discussion of the topic and points out that understanding those issues involves activities unfamiliar to economists but concludes that "messy or not, they are important and must be pursued".

Frey (1997) introduces a principal-agent model where he allows for the possi-

[3] The expression "motivation crowding-out" has been coined by economist Frey (1997). A description of cognitive evaluation theory is for instance given by Deci and Porac (1978). Self-perception approaches building on the work by Bem (1967) are described by Lepper and Greene (1978).

[4] Frey and Jegen (2001) or Kunz and Pfaff (2002) review the results of the psychological experiments from an economic perspective.

[5] Incentives are set with a bonus or a fine in Fehr and Gächter (2002) and with a piece rate in Irlenbusch and Sliwka (2003).

[6] Related with this effect but not concerned with effort for a task is Frey and Oberholzer-Gee (1997). In this study, questionaire data is analyzed indicating that people's willingness to accept that a nuclear waste disposal facility in their neighbourhood is lower when monetary compensation is offered for the acceptance.

bility that an agent's disutility of effort is increasing in the monetary reward for this effort. Hence, he is able to show that principal-agent theory can in principle be adapted to encompass the reduction of intrinsic motivation by monetary rewards. However, the crowding-out effect, i.e. the reason for the positive relationship between disutility of effort and rewards, is not shown endogenously in the model. This is of course legitimate as one can refer to psychological results and theories to justify such an assumption and analyze its effects in a standard framework. However, it seems also interesting to consider whether an explanation for motivation crowding-out can be given endogenously within the economic model. In this paper we want to follow the latter approach.

To do this we extend a simple standard principal-agent model in two steps. As a first step, we introduce the possibility that a certain fraction of agents in the population is *reliable*, i.e. such agents stick to agreements they signed even if the fulfillment is not verifiable.[7] If this is the case, a contract consists – besides the payment scheme – also of a performance objective specifying the effort level the agent should choose. In contrast to a "homo oeconomicus", a reliable agent sticks to such an agreement once he has signed the contract.

We derive optimal incentive contracts for this case. If performance measurement is costly, we show that incentive contracts are optimal if and only if the fraction of reliable agents is sufficiently small and otherwise fixed wage contracts are chosen. Furthermore, as long as the fraction of reliable agents is positive, optimal incentive contracts are lower powered than standard theory predicts. In this view, the purpose of incentive contracts is *to insure the principal against contracting with unreliable agents*. Incentive contracts are completely useless if all agents are reliable as they then only impose wasteful cost for risk premia and performance measurement.

In a second step, we consider the importance of *social norms* in the following way: We depart from the idea that the "moral convictions" of all agents, i.e. their attitude on sticking to agreements, are fixed in advance. Instead we assume that a third group of agents exists, who do not fall yet into one of the categories of being either reliable or unreliable but are influenced in their moral convictions by what they think others will do. We assume that those "conformists" are reliable

[7] Koford and Penno (1992) dicuss the possibility of contracting with ethical agents who do not cheat from an accounting perspective. They show for instance that the installation of internal control systems becomes less profitable if the proportion of ethical agents rises.

if the social norm “requires” to be reliable. More precisely, a confomist will be reliable if and only if she or he thinks that a sufficiently high fraction of the other “steadfast” agents is reliable as well. The game we then analyze is therefore in a very simple way a *psychological game* as defined by Geanakoplos et al. (1989), where players’ payoffs may not only be affected by what they do but also by what they believe.

Typically, in reality there will be some uncertainty on the prevalence of a social norm. Translated to our model, the fraction of reliable agents in the population will not perfectly known by an individual agent. The principal as the employer, however, is supposed to have superior information on the prevalent norm in form of a more precise signal about the fraction of reliable agents for instance as she has experienced the behavior of previous employees.

If now the principal chooses the compensation scheme based on this superior information, the agents learn something on the principal’s beliefs about the prevailing norm. In particular, an agent may conclude from being offered a high powered incentive scheme that the principal thinks that the fraction of reliable agents is small: Otherwise, she would not have chosen the costly incentive contract. But in that case, the agent learns that not sticking to agreements is a widespread and therefore acceptable behavior. Hence, a conformist who is influenced by this information may become unreliable after being offered an incentive contract whereas the same agent might have stayed reliable with a pure fixed wage. As we will show, given any optimal contract a reliable agent always exerts a higher effort than an unreliable agent. Hence, incentive contracts may indeed reduce the motivation of agents. A principal who does not take that into account ignores those ‘*hidden costs of incentive contracts*’ and may actually be worse off.

A rational principal will of course take these indirect consequences into account. She has to trade off two effects when designing optimal compensation schemes. On the one hand, there is the well known incentive effect: Higher powered incentives lead to higher effort levels of unreliable agents. But on the other hand, a crowding-out effect may arise as the choice of high powered incentives may signal that not being reliable is a widespread and acceptable demeanor, which may undermine the reliability of others. It is then of course interesting to know whether she will still choose incentive contracts or prefer a fixed wage to avoid motivation crowding-out. We will therefore analyze the circumstances under which the latter effect outweighs the former and will characterize equilibria of the game arising.

A different approach to explain motivation crowding-out has very recently been taken by Bénabou and Tirole (2002), who formalize aspects of psychological self-perception approaches. A key assumption in their principal-agent model is that an agent does not know perfectly whether he likes or dislikes a task he has to fulfill. But the principal has more accurate information on whether the task is attractive for the agent in the end. Motivation crowding-out takes place in their model, as the agent infers from being offered a high powered incentive scheme that the principal thinks that he will dislike the task, which in turn reduces his own beliefs on the attractiveness of the task. A difference between our approach and Bénabou and Tirole's as well as the mentioned psychological explanations for motivation crowding-out is that a necessary precondition for those explanations is that an agent likes performing the task with positive probability, whereas our approach can explain motivation crowding-out also in those circumstances typically analyzed in principal-agent models where an agent always dislikes higher efforts.[8]

The paper proceeds as follows. Section 2 presents the simple principal-agent model. In Section 3 optimal compensation schemes are analyzed in this model. Conformist agents who are influenced by social norms are introduced in Section 4 and the consequences for optimal contract offers are analyzed in this section. Finally, Section 5 concludes.

2 A Simple Model

Our model simply extends a standard Holmström/Milgrom or LEN-type linear principal-agent model with normally distributed noise used in numerous applications[9] to encompass the possibility of contracting with reliable agents. A principal P employs an agent A to perform a certain task. The agent can exert an effort

[8] Note that the experiments by Fehr and Gächter (2002) and Irlenbusch and Sliwka (2003) mentioned above have found that incentives reduced motivation in situations where the agents unambiguously disliked higher effort as higher effort simply was represented by a higher monetary expenditure by the agent.

[9] Compare Holmström and Milgrom (1990), Holmström and Milgrom (1991), Gibbons and Murphy (1992), Holmström and Milgrom (1994) or in accounting for instance Feltham and Xie (1994), Datar et al. (2001) and many more. For an overview see also Prendergast (1999).

level e at costs $c(e)$ where $c(e) = \frac{c}{2}e^2$. The effort level e is unverifiable. In a first step, we assume that there are two types of agents, reliable and unreliable ones. The agent's type is denoted by θ. If $\theta = \theta_R$ the agent is reliable, if $\theta = \theta_U$ he is unreliable. All agents maximize their individual utility but reliable agents stick to promises or agreements they made. One can think of this in the following simple way: Once a reliable agent has promised to choose a certain action he feels guilty when not fulfilling this promise and will receive an extremely large utility loss from having a bad conscience such that it is always in his best interest to stick to the promise. The principal does not know whether the agent is reliable or unreliable but she has a prior belief that the agent is reliable with probability ϕ. In Section 4 we will generalize this to encompass the possibility that agents can still be influenced in their moral convictions.

The principal's revenue from the agent's task simply corresponds to the agents effort level e. This payoff is itself unverifiable. In contrast to standard hidden action models we assume that performance measurement is costly: At a given cost k the principal can install a technology yielding a verifiable performance signal x. This assumption should capture the fact that typically in practice the installation of a performance measurement system and the payout of variable wage components come at a cost for the company. The performance signal x is given by

$$x = e + \varepsilon,$$

where ε is a normally distributed random variable with mean 0 and variance σ^2. If the principal decided to set up an incentive system, she can offer a linear compensation scheme consisting of a fixed wage α and a variable rate β such that the agent receives a total wage of

$$\alpha + \beta(e + \varepsilon)$$

when choosing an effort level of e. But note that the principal has an additional contractual instrument: She can specify a requested effort level or performance commitment in the contract, which we denote by $\hat{e}$. Once a reliable agent has signed such a contract he will honor this performance commitment and choose $e = \hat{e}$ even though the effort level is unverifiable. We can denote any such contract by a vector $C = (\hat{e}, \alpha, \beta)$.

If the principal decides against such an incentive system, she saves the costs of performance measurement k and contracts simply consist of a fixed wage α and a requested effort level of $\hat{e}$.

We assume that the principal can only propose a single contract at a time and therefore cannot screen the agents.[10] The principal is risk neutral and the agent is risk averse with constant absolute risk aversion. An unreliable agent's utility function when receiving a wage payment w and choosing e is given by

$$-\exp\left(-r\left(w-c\left(e\right)\right)\right),$$

where r is the agent's rate of absolute risk aversion. The utility function of the reliable agent additionally depends on an agreed effort $\hat{e}$. If $e=\hat{e}$ he has the same utility function as an unreliable agent. If $e\neq\hat{e}$ his utility is minus infinity.[11] All agents have a reservation wage of 0.

3 Optimal Compensation Schemes

We start by analyzing the optimal compensation when the principal selects a pure fixed wage given that she knows that the agent is reliable with probability ϕ. Then we proceed by investigating optimal incentive schemes for that case, compare both types of compensation and discuss their relative advantages.

3.1 Fixed Wage Contracts

If the principal intends to set a fixed wage, a contract offer is simply $C=(\hat{e},\alpha,0)$. An unreliable agent will then of course never exert any effort even if he accepted the contract. A reliable agent will accept the contract if it is in his best interest, i.e. if his participation constraint

$$\alpha-c\left(\hat{e}\right)\geq 0$$

[10] See Alger and Renault (2000) for a paper where optimal screening is analyzed in an adverse selection framework without hidden actions in which some agents are honest.

[11] The preferences are therefore lexicographic: The utility of any bundle $(e,\hat{e},w)$ where $e\neq\hat{e}$ is lower than the utility of any bundle $(e,\hat{e},w)$ with $e=\hat{e}$.

is met and will afterwards always choose $\hat{e}$. This participation constraint will always be binding as the principal could otherwise reduce α and would be strictly better off. Therefore, the wage will compensate for the reliable agents' effort costs and, hence, $\alpha = c(\hat{e})$. The principal maximizes her expected payoff by taking this into account

$$\max_{\hat{e}} \phi\hat{e} - c(\hat{e}).$$

The optimal requested effort level is defined by $\phi = c'(\hat{e})$. Hence, as costs are quadratic we get that $\hat{e} = \phi/c$ and the optimal wage payment is $\alpha = \phi^2/2c$, which gives us the following simple result for the optimal fixed wage contract denoted by $C_F^*(\phi)$:

Proposition 1 *The optimal contract $C_F^*(\phi) = (\hat{e}, \alpha, 0)$ in a pure fixed wage regime is defined by*

$$C_F^*(\phi) = \left(\frac{\phi}{c}, \frac{\phi^2}{2c}, 0\right).$$

Note that an unreliable agent earns a rent of size α. The principal will offer a positive wage level as long as $\phi > 0$ as she earns strictly positive profits from the reliable agents. The higher the effort $\hat{e}$ the principal requires from the reliable agents the higher has to be the wage to compensate them for their costs of effort. But the larger will also be the loss when the agent turns out to be unreliable. The larger is the fraction of reliable agents in the population, the larger the expected return for a given wage payment. Hence, the wages and the requested effort levels increase with the fraction ϕ of reliable agents.

If all agents are reliable (i.e. $\phi = 1$) the first-best solution is attained as a contract is chosen where the marginal costs of the requested effort are equal to the marginal return to the principal which is equal to 1. The principal's first-best profit is therefore $1/2c$.

A huge and steadily growing experimental literature[12] has examined such "gift exchange" settings based on Akerlof's (1982) fair wage-effort hypothesis in which a player in the role of a principal can offer a wage w to another player in the role of an agent, who in turn can choose an effort level e which is costly for the agent but beneficial for the principal.[13] Whereas standard theory (i.e. $\phi = 0$ in our model)

[12] Compare for instance Fehr et al. (1993), Berg et al. (1995), Fehr et al. (1997) or for a recent overview Fehr and Schmidt (2002).

[13] In those experiments the agent typically chooses a number representing his effort level. The higher the number, the higher are the profits for the principal but also the costs for the agent.

predicts that $w = e = 0$ in that case, in many experiments it has been shown that players typically choose positive wages and effort levels. Moreover, on average the effort exerted by an agent increased in the size of the wage he received. Mainly, this has been attributed to preferences for reciprocity or inequity aversion of the players.[14] Although we do not doubt that reciprocity and inequity aversion play an important role, this model yields a different simple possible explanation for the monotonic relationship of efforts and wages: High requested effort levels go along with high wage offers and reliable agents will then indeed respond with high effort levels. Indeed, most of the studies examine situations as considered here, where the principal proposes a wage in connection with a requested effort level before the agent chooses his actual effort, although for theories of inequity aversion and reciprocity this effort request is neither necessary nor should have any impact on the outcome.

3.2 Incentive Contracts

Once the principal has chosen to set up a performance measurement system at costs k she can propose a performance contingent incentive contract and therefore has one additional instrument to steer the agent's actions. Suppose the principal offers a contract $C = (\hat{e}, \alpha, \beta)$ with a positive value of β. If the agent is reliable, he will always choose $\hat{e}$ after having accepted the contract. But we now have to take into account that even though β does not affect this reliable agent's motivation it is of course of importance for his expected utility from accepting the contract: On the one hand, he receives an additional share of the profits. On the other hand, his income will become risky as the performance measure is noisy. Again, the contract must yield at least as much utility as his reservation wage 0. Hence, a reliable agent will accept the contract and choose $\hat{e}$ if

$$\alpha + \beta\hat{e} - \frac{c}{2}\hat{e}^2 - \frac{1}{2}r\sigma^2\beta^2 \geq 0. \tag{1}$$

Note that the same contract will always be accepted by an unreliable agent if the inequality is met: He is at least weakly better off than the reliable agent as he may

[14] Reciprocity has been formally modelled by Rabin (1993), Falk and Fischbacher (1999) or Dufwenberg and Kirchsteiger (1999), inequity aversion by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). A more general approach integrating both forms of social preferences is introduced by Charness and Rabin (2002).

always choose $\hat{e}$ as well. Hence, there is no contract offer that is accepted only by reliable agents. Any contract satisfying (1) will be accepted by the unreliable agent and the latter will afterwards choose e by simply maximizing his utility. The first order condition of his objective function yields

$$\beta - c'(e) = 0 \Leftrightarrow e = \frac{\beta}{c} \tag{2}$$

as costs are quadratic. Note that the principal is of course at least weakly better off when contracting with a reliable agent as she can always set $\hat{e} = \beta/c$. Hence, she clearly always wants that reliable agents accept the contract. But as we have seen, she cannot prevent unreliable agents from accepting.

We now denote by e the incentive compatible effort level chosen by the unreliable agents in contrast to the effort level $\hat{e}$ chosen by the reliable agents. The principal's expected payoff when offering an incentive contract $(\alpha, \beta, \hat{e})$ is then given by

$$(1-\beta)\left[\phi\hat{e} + (1-\phi)\, e\right] - \alpha - k.$$

The optimal contract choice amounts to the maximization of the principal's expected payoff subject to the incentive constraint of the unreliable agent (2) and the participation constraint of a reliable agent (1) which implies that of an unreliable agent:

$$\begin{aligned} \max_{\alpha,\beta,\hat{e},e} \quad & (1-\beta)\left[\phi\hat{e} + (1-\phi)\, e\right] - \alpha - k \\ \text{s.t.} \quad & \alpha + \beta\hat{e} - \frac{c}{2}\hat{e}^2 - \frac{1}{2}r\sigma^2\beta^2 \geq 0 \\ & \beta - ce = 0 \end{aligned}$$

As in standard hidden action problems with transferable utility and unlimited liability the participation constraint will be binding as otherwise the fixed wage α could be reduced. Hence, we can solve the binding condition (1) for α and substitute this in the principal's objective function. By solving the program we obtain the following result for the optimal incentive contract denoted by $C_V^*(\phi)$:

Proposition 2 *The optimal incentive contract $C_V^*(\phi) = (\hat{e}, \alpha, \beta)$ has the follow-*

ing properties: The variable rate is

$$\beta = \frac{1-\phi^2}{1-\phi^2+cr\sigma^2}.$$

The smaller the fraction of reliable agents the higher powered is the incentive scheme. The performance objective $\hat{e}$ is larger than the incentive compatible effort level e, where

$$\hat{e} = \frac{1-\phi^2+\phi cr\sigma^2}{c\left(1-\phi^2+cr\sigma^2\right)} \text{ and } e = \frac{1-\phi^2}{c\left(1-\phi^2+cr\sigma^2\right)}.$$

Proof: See Appendix.

Note that for $\phi = 0$ the solution is exactly the standard second best solution in the linear contracts model. The higher the fraction of reliable agents, the lower powered is the incentive scheme. For $\phi = 1$ the variable rate β is zero and the contract is a pure fixed wage contract.

A direct and important implication is that incentive contracts are completely unnecessary if all agents are reliable. If that is the case, choosing an incentive contract only imposes a wasteful risk on the risk averse agent and the principal has to bear the direct costs of measuring performance. However, the higher the danger of contracting with an unreliable agent, the higher powered the incentive scheme.

Hence, in the spirit of our model, the function of an incentive contract is to insure the principal against the *risk of contracting with unreliable agents.* The principal chooses higher powered incentive schemes only if she fears that the work ethics of her employees is bad.

In this sense, the simple model captures the notion of 'moral hazard' in a more general way than standard agency theory does. The term 'moral hazard' originated as the "risk to an insurance company resulting from uncertainty about the honesty of the insured"[15]. But in standard moral hazard models there is no uncertainty about the honesty of agents, as they are always dishonest. Our model

[15] Compare The American Heritage Dictionary of the English Language: Fourth Edition. 2000.

indicates that those models predict too high powered incentives if there is indeed some uncertainty about the honesty of agents.

3.3 Fixed Wages or Pay-for-Performance?

Now we compare both contract types. The key question is whether the principal wants to set up a costly performance measurement system. As we have seen, she will never need to set-up an incentive system if all agents are reliable (i.e. $\phi = 1$). In that case, the first best is attained by a simple fixed wage contract. On the other hand, if all agents are unreliable ($\phi = 0$), we are back in the standard second best world: The principal cannot attain a positive payoff without a performance contingent contract.

It is useful to introduce some additional notation. We denote by $C^* (\phi)$ the optimal (fixed wage or incentive) contract given that the probability of meeting a reliable agent is ϕ. The principal's profit for a given contract C and probability ϕ is denoted by $\pi (C|\phi)$. We now compare the profits in both systems and obtain the following result:

Proposition 3 *The principal will choose an incentive contract if and only if the fraction of reliable agents is smaller than a cut-off value $\hat{\phi}$ and otherwise a fixed wage. The principal's profit is*

$$\pi (C^* (\phi) |\phi) = \begin{cases} \frac{\phi^2}{2c} & \text{if } \phi \geq \hat{\phi}, \\ \frac{1}{2c} - \frac{\left(1-\phi^2\right) r\sigma^2}{2\left(1-\phi^2+cr\sigma^2\right)} - k & \text{otherwise.} \end{cases} \tag{3}$$

The cut-off value $\hat{\phi}$ is decreasing in the agents' risk aversion r, the variance of the performance measure σ^2 and the costs of the incentive scheme k. The profit function is strictly increasing in the proportion of reliable agents ϕ.

Proof: See Appendix.

With a pure fixed wage as well as with an incentive scheme, the profit function is strictly increasing in the fraction of reliable agents ϕ. Hence, the same holds for the composite function. An example for the profit function is plotted in Figure 1. For values of ϕ smaller than the cut-off $\hat{\phi}$ it is determined by the profits with an optimal incentive scheme. For values larger than $\hat{\phi}$ the principal will set a fixed wage. When $\phi = 1$, the first-best solution is attained with a fixed wage. The value of the profit function at $\phi = 0$ is just determined by the principal's profit with a standard optimal incentive contract.

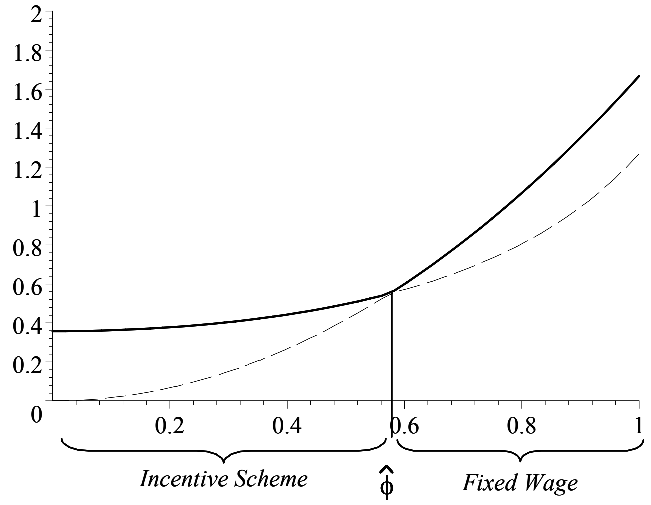

Figure 1: The principal's profit as a function of the fraction of honest agents ϕ.

4 Contracts Signal Social Norms

So far we have assumed, that it is clear for an agent from the outset that he is either reliable or unreliable. One may think of this as a situation where there is the social norm: "you have to stick to your word" and some agents adhere to this norm while others ignore promises and seek to maximize individual monetary payoffs. Social psychologists define social norms as people's beliefs about the attitudes and behaviors that are normal or acceptable in a particular social context. In many situations, people's perception of these norms will greatly influence their behavior. Hence, when it is a question of social norms whether it is acceptable to be unreliable, the behavior of people is essentially affected by their beliefs on the behavior of others. We now extend our simple model to encompass such a situation.

First, we differentiate between agents who have a firm disposition and others who don't. A fraction η of all agents – called the *steadfast* agents – is assumed to have already adopted a norm as those considered in the previous section. For them it is clear from the beginning that they have a reliable or unreliable disposition. As before, the reliable agents among the steadfast have a utility function, such that they suffer an infinite utility loss if they fail to choose $\hat{e}$ once having signed the contract whereas an unreliable agent's utility function is independent of $\hat{e}$.

The other fraction $1 - \eta$ of agents, however, has not "made up their mind" which norm to follow. We assume that those agents – we will call them *conformists* – will be reliable, if and only if they believe that the fraction of reliable agents among those who have a firm disposition is sufficiently large. Note that the process of adopting a norm by the conformists is therefore no conscious rational decision. Hence, we depart from a standard game theoretic framework and assume that the agent's preferences are affected by their beliefs on the preferences of others. Therefore the game we analyze is in a very simple way a *psychological game* in the sense of Geanakoplos et al. (1989), where a player's payoff does not only depend on what he and others do but also on what he and others think.

Our intention is simply to capture the idea that some people tend to stick to their word if they think that others also keep their promises but fail to do so, if they think that most others do not stick to promises either. Again, the agents type is denoted by θ. Now, an agent can be of three different types. First, there are steadfast agents who already adhere to a norm and are always reliable ($\theta = \theta_R$) then there are those who are steadfast but unreliable ($\theta = \theta_U$) and finally there are conformists ($\theta = \theta_C$).

But people are typically not perfectly informed about the moral principles of others and there is uncertainty on the predominant social norm. We therefore assume, that the fraction of reliable agents is only imperfectly known to principal as well as agents. Both, the principal and all agents are assumed to share a common prior belief on the fraction of reliable agents among the steadfast denoted by τ, which is drawn from some prior distribution.

But, typically an employer will have more information on the reliability and work habits of his employees than for instance an agent who is new to a particular firm or organization from past experience with other employees. To capture this idea, we suppose that before making a contract offer to the agent the principal receives an additional signal $s \in \{L, H\}$ on the fraction of reliable agents among the steadfast. We denote by

$$\tau_s = \Pr\{\textit{Agent reliable}|\textit{Agent steadfast},\ \textit{Signal } s\} = E[\tau|s]$$

the conditional expectation on τ given signal s. A signal $s = L$ indicates a low fraction of reliable agents and a signal H a high fraction. Hence, we assume that $\tau_H > \tau_L$. This signal is not observed by the agent.

It remains to describe under which conditions a conformist turns out to be reliable or unreliable. In this respect, we assume that an agent who has not yet adopted a norm will behave reliably if he believes that τ is higher than a given cut-off value $\bar{\tau}$ in expected terms.

The timing of the extended model is as follows:

1	2	3	4	5
Nature determines τ and agent's type θ	Agent learns type θ; Principal learns signal s	Principal proposes contract $(\hat{e}, \alpha, \beta)$	Agent can accept or reject $d \in \{y, n\}$	Agent chooses effort e

Hence, after the fraction τ of reliable agents is determined, the type of the agent is chosen. With probability $\eta\tau$ the agent is always reliable, with probability $\eta(1-\tau)$ he is always unreliable and with probability $1-\eta$ he is a conformist. The agent learns his type and the principal receives a signal on the fraction of reliable agents. After that, the principal makes a contract offer to the agent. The agent then updates his beliefs on the fraction of reliable agents and decides on whether to accept the contract ($d = y$) or reject it. After that the agent chooses the effort level. If he is either reliable right away or a "conformist converted to reliability" he will always choose $\hat{e}$ if he signed the contract. Finally, the output x is realized as above and the agent receives a wage payment as specified in the contract.

4.1 Motivation Crowding-Out

The principal has superior information on the fraction of reliable agents in the population when making the contract offer. We might thus expect that her contract choice is influenced by this information. As we have seen in the preceding section, she tends to choose a higher powered incentive scheme if the fraction of reliable agents is small. On the other hand, it is attractive to choose a fixed wage if the fraction of reliable agents is large.

But of course, the choice of the wage scheme conveys information to the agent receiving the offer. As the principal has superior information, the contract choice will strongly influence the agent's beliefs on the prevalent norm. Suppose that the principal indeed makes different contract offers depending on her signal about the fraction of reliable agents. In that case, a conformist can learn from being offered a higher powered incentive scheme that the principal thinks that not sticking to promises is a widespread and therefore "acceptable" demeanor. But this may lead a conformist to become unreliable as he may be lead to choose only the incentive compatible effort level instead of the higher performance objective agreed on in the contract.

From these considerations it becomes evident that choosing higher powered incentives may in principle have two effects on the agent's motivation:

- A well known *incentive effect*: Unreliable agents have higher incentives to exert effort simply by striving to maximize their material payoff.
- A *motivation crowding-out effect*: Offering higher powered incentive schemes signals that not sticking to an unverifiable performance objective is a widespread and hence acceptable behavior and this makes conformists to become unreliable.

The game described above is a signalling game. Of course, a rational principal will anticipate both the incentive as well as the crowding-out effect. It is therefore interesting to analyze whether crowding-out may indeed take place in equilibrium. We seek for Perfect Bayesian Equilibria in pure strategies of this game, which consist of: the principal's strategy as a contract offer $C_s = (\hat{e}_s, \alpha_s, \beta_s)$ for each information set $s \in \{L, H\}$, an agent's strategy consisting of an acceptance decision $d(C, \theta)$ and effort choice $e(C, \theta)$ for each contract offer C and type θ

and finally, an agent's beliefs $\Pr\{s = H|C, \theta\}$ on the principal's signal for each contract offer C and type θ.

We start by analyzing how a conformist's reliability is determined, given his beliefs on the principal's signal on the equilibrium path. If the principal's private information on the norm is not revealed, the agent's beliefs are unaffected. Hence, a conformist will a priori be reliable if and only if the agents' prior beliefs are such that $E[\tau] \geq \bar{\tau}$.

If, however, the agent learns the principal's information, this signal may only have an impact on the agent's reliability if $\tau_H > \bar{\tau} > \tau_L$. Therefore we impose this condition as an assumption as otherwise the conformists will either always be reliable or unreliable, in which case the analysis of section 3 applies:

Assumption 1: $\tau_H > \bar{\tau} > \tau_L$.

A conformist will then be reliable if he learns that the principal received a high signal and unreliable if he learns that it has been a low signal.

The principal's contract choice will of course depend on her beliefs about the agent's reliability given her information and the agent's behavior. Her interim belief that the agent is reliable is determined as follows:

$$(1-\eta)\Pr\{Conformist\ is\ reliable\} + \eta E[\tau|s].$$

We denote the overall probability that the principal meets a reliable agent if all conformists are reliable (R) and she received a signal that the fraction of reliable agents is high (H) by ϕ_{RH}, where

$$\phi_{RH} = (1-\eta) + \eta\tau_H.$$

If however, the principal received a low signal and this is learned by the agent any conformist will be unreliable and the overall probability of contracting with a reliable agent is[16]

$$\phi_{UL} = \eta\tau_L.$$

Recall that a cut-off level $\hat{\phi}$ was defined in Proposition 3, such that it is in

[16] Similarly we define $\phi_{UH} = \eta\tau_H$ and $\phi_{RL} = (1-\eta) + \eta\tau_L$.

the principal's best interest to choose a fixed wage contract if and only if the probability of contracting with a reliable agent is larger than this cut-off. We now impose the assumption that the principal prefers to choose a fixed wage contract when all agents are steadfast and the signal indicates that a high fraction of them is reliable. On the other hand, we suppose that she would rather choose an incentive contract when again all are steadfast but only a low fraction is reliable:

Assumption 2: $\tau_H > \hat{\phi} > \tau_L$.

In particular, this assumption implies that $\phi_{RH} > \hat{\phi} > \phi_{UL}$ for all values of η.[17] Hence, the principal always prefers a fixed wage contract when all conformists and a high fraction of the steadfast agents are reliable but rather chooses an incentive contract if the conformists are unreliable and only a low fraction of steadfast is reliable. Assumptions 1 and 2 assure that potentially fixed wages as well as incentive schemes can in principle be optimal when the principal's information is revealed.

4.2 Equilibrium Analysis

Now we can analyze, whether motivation crowding-out may indeed take place in equilibrium. Note that the principal may for instance have an interest not to choose a high powered incentive scheme when her signal is low but to stick to the same contract she would offer after the high signal. By this she might be able to prevent the crowding-out effect among the conformists, even if incentives for the steadfastly unreliable agents are forgone. If this is the case, a pooling equilibrium arises in which the principal offers the same contract independent of her signal.

4.2.1 Pooling Equilibria: Preventing Motivation Crowding-Out

However, a precondition for the existence of a pooling equilibrium is that the conformists are reliable a priori in the sense that they initially believe that sticking to an agreement is the prevailing norm ($E[\tau] \geq \bar{\tau}$). To see this, just suppose the contrary holds and the conformists a priori believe that less than a fraction $\bar{\tau}$ of the steadfast stick to a norm in expected terms. In that case, any conformist

[17] To see that, note that these inequalities always hold if $\eta = 1$ as then they correspond to the assumption. Furthermore, ϕ_{RH} is decreasing and ϕ_{UL} increasing in η.

will behave unreliably when no further information is revealed. As in a pooling equilibrium no information is revealed, any conformist will be unreliable in equilibrium. But then motivation crowding-out cannot take place if the principal deviates from the pooling contract, as the motivation to stick to agreements is already absent among the conformists. The principal is then best off when choosing the optimal contract given her signal. As we have seen in the previous section she will optimally propose different contracts in both states and this in turn will reveal her information.

However, when the conformists are reliable a-priori then pooling equilibria can arise as the following result shows:[18]

Proposition 4 *(i) If* $E[\tau] < \bar{\tau}$ *there is no pooling equilibrium. (ii) If* $E[\tau] \geq \bar{\tau}$ *a pooling equilibrium exists in which a fixed wage contract is offered regardless of the principal's signal if and only if this contract* $C_P = (\hat{e}_P, \alpha_P, 0)$ *satisfies the following conditions*

$$\begin{aligned} \alpha_P &\leq \phi_{RH}\hat{e}_P - \pi\left(C^*\left(\phi_{UH}\right)|\phi_{UH}\right), & (4)\\ \alpha_P &\leq \phi_{RL}\hat{e}_P - \pi\left(C^*\left(\phi_{UL}\right)|\phi_{UL}\right), & (5)\\ \alpha_P &\geq \frac{c}{2}\hat{e}_P^2. & (6) \end{aligned}$$

(iii) The set of fixed wage pooling equilibria is non-empty if and only if the fraction of steadfast agents η *is smaller than a cut-off level.*

Proof: See Appendix.

Note that if any pooling equilibrium can be sustained, then it can also be sustained with out-of-equilibrium beliefs such that the agent thinks that the principal received a low signal after any deviation from the equilibrium path. In that case, any deviation will lead the conformists to become unreliable and – as we have seen in Proposition 3 – the principal is worse off with a higher fraction of unreliable agents. A fixed wage pooling equilibrium in which a contract C_P is offered exists if the principal has no incentive to deviate from C_P whatever her signal. After having received the high signal, the highest possible expected payoff she

[18] Recall the notation from the previous section: $\pi(C|\phi)$ is the principal's expected profit from a contract C if the fraction of reliable agents is ϕ. $C^*(\phi)$ is the optimal contract when the fraction of reliable agents is ϕ.

can receive when all conformists are unreliable is $\pi\left(C^{*}\left(\phi_{UH}\right)|\phi_{UH}\right)$ as given by Proposition 3. Condition (4) guarantees that the principal is better off with C_P in that case. Condition (5) does the same for the case of a low signal. Finally, condition (6) is the agent's participation constraint.

A pooling equilibrium with a pure fixed wage exists if a contract C_P satisfies all three conditions. As Proposition 4 has shown this will be the case if and only if the prior is such that all conformists are initially reliable and the fraction of conformists is large enough.

To understand this on an intuitive level, it is important to note that the reason why the principal sticks to the pooling contract in equilibrium is that she wants to prevent motivation crowding-out. If, for instance, she has received a signal indicating a low proportion of reliable agents and therefore proposes an incentive contract this will lead the conformists to become unreliable. If the fraction of conformists is large (i.e. η is small), this effect will be of high importance as the probability that the agent's behavior is affected is large. In addition, the proportion of steadfast agents is small and, as a consequence, there are also only a few steadfastly unreliable agents who are positively affected by incentives. The *motivation crowding-out effect* may therefore outweigh the *incentive effect*.

If however, the number of conformists is relatively small and therefore the number of steadfast agents large, the crowding-out effect will be less important. In that case, it is more important to fine tune the incentive scheme according to the signal and the principal has to worry less about the reduction of work ethics. A pooling equilibrium will therefore never occur if there are only a few conformists or – putting it differently – if the uncertainty on the behavioral norm is small.

Figure 2 shows possible fixed wage pooling contracts for an example.[19] The requested effort levels $\hat{e}$ are plotted on the abscissa, the wages α on the ordinate axis. The upper ascending line is the upper boundary of the half-plane defined by (4): Offering a fixed wage above this line would not be attractive for a principal having received a high signal. Wages are so high that she would be better off by offering a different contract even if the conformists get unreliable. Similarly, the lower ascending line is the upper boundary of the half-plane (5).The set of fixed wage contracts sustainable in a pooling equilibrium is in the darkly shaded area

[19] The parameter values are $c = 1.1$, $k = 0.3$, $\eta = 0.5$, $\tau_L = 0.9$, $\tau_S = 0.3$, $r = 0.2$ and $\sigma = 1$.

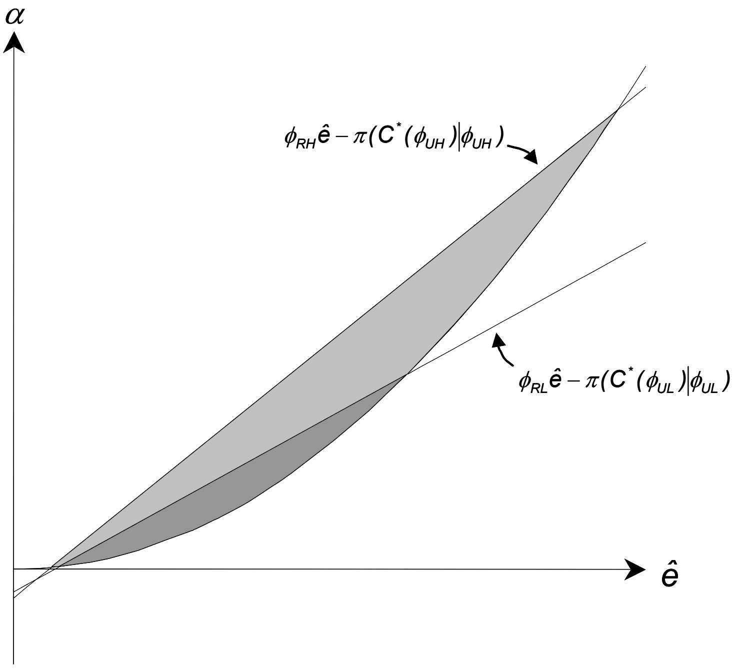
α
$\phi_{RH}\hat{e} - \pi(C^*(\phi_{UH})|\phi_{UH})$
$\phi_{RL}\hat{e} - \pi(C^*(\phi_{UL})|\phi_{UL})$
$\hat{e}$

Figure 2: Equilibrium fixed wage offers

below these two boundaries where both constraints are met and above the cost function.

4.2.2 Separating Equilibria

It is now of course interesting whether separating equilibria exist. Note that the principal's contract offer then reveals her private information on the prevailing norm. She offers a different contract when she thinks that the fraction of reliable agents among the steadfast is high than when it is low. This of course has the direct consequences for the conformists' moral disposition laid out above. Hence, in any separating equilibrium motivation crowding-out will indeed take place.

We focus on those equilibria in which the principal offers a fixed wage contract after having observed a signal indicating a high fraction of reliable agents. Indeed, we can show that such separating equilibria always exist:

Proposition 5 *There is a set of separating equilibria with the following properties: (i)The principal proposes the incentive contract $C_L = C^* (\phi_{UL})$ after having observed the low signal $s = L$. (ii) The principal offers a fixed wage contract*

$C_H = (\hat{e}_H, \alpha_H, 0)$ *after having observed the high signal. This contract must have the following properties:*

$$\alpha_H \leq \phi_{RH}\hat{e}_H - \pi\left(C^*\left(\phi_{UH}\right)|\phi_{UH}\right), \tag{7}$$

$$\alpha_H \geq \phi_{RL}\hat{e}_H - \pi\left(C^*\left(\phi_{UL}\right)|\phi_{UL}\right), \tag{8}$$

$$\alpha_H \geq \frac{c}{2}\hat{e}_H^2. \tag{9}$$

(iii) This set is always non-empty.

Proof: See Appendix.

A separating equilibrium consists of two contracts. The contract C_L is offered after the principal received the low signal and C_H after the high signal. On the one hand, if a separating equilibrium exists, neither the principal should have an incentive to pretend having received the high signal (by offering C_H) when she received the low signal nor vice versa. But in addition, in both cases the principal should have no interest to deviate to a contract different from C_L and C_H. If any separating equilibrium can be sustained, then there will be one where agents believe that the principal received the low signal after any deviation from the equilibrium path. It can be straightforwardly seen that only $C^*\left(\phi_{UL}\right)$ can be chosen after the low signal in a separating equilibrium: When choosing C_L the principal reveals that she received the low signal. Any deviation (except that to C_H) will also make them think that she received the low signal. But $C^*\left(\phi_{UL}\right)$ is the best contract choice for that case. Still, we have to guarantee that the principal has no incentive to deviate and choose C_H after the low signal which will make the agents think that she believes in a high fraction of reliable agents. But this is ensured by condition (8).

When the principal received the high signal, any deviation from C_H will lead the conformists to become unreliable (including of course a deviation to C_L). The best possible deviation is to choose $C^*\left(\phi_{UH}\right)$. Hence, for a fixed wage contract C_H to be part of an equilibrium the principal must be better off with C_H in the case where all conformists are reliable than with $C^*\left(\phi_{UH}\right)$ with unreliable conformists. This condition is equivalent to condition (7). In addition, the contract C_H must satisfy the participation constraint (9). As we have shown, there are always fixed wage contracts C_H that satisfy those constraints and therefore the set of separating equilibria is always non-empty.

The possible fixed wage contracts offered in the separating equilibrium after a high signal are also illustrated in figure 2 in the lightly shaded area. Note that the boundaries coincide with those given for the pooling equilibrum with fixed wages. But here, the lower ascending line is the *lower* boundary of the half-plane (8): Contracts below this line would be imitated by a principal having received a low signal and therefore cannot be part of a separating equilibrium (this was a precondition for such a contract to be part of a pooling equilibrium). The upper ascending line is again the upper boundary of the half-plane (7): Offering a fixed wage above this line would not be attractive for a principal having received a high signal. Wages are so high that she would be better off by offering a different contract even if the conformists get unreliable.

Figure 3 shows another example where there are less conformists in the population.[20] The fixed wage contracts in the lightly shaded area satisfy both constraints (8) and (7) as well as the participation constraint. Therefore, separating equilibria exist in which such contracts are offered by the principal after she received a high signal. However, the lower boundary for (8) which coincides with the upper boundary for (5) in the pooling case is here always below the cost function. Any fixed wage contract satisfying the participation constraint is worse for a principal with a low signal than her optimal incentive contract. Hence, fixed wage pooling equilibria cannot exist.

When the principal received a signal indicating that the fraction of reliable agents in the population is small, she will offer an incentive contract in such a separating equilibrium. By doing this she reveals the information that a large part of the population is not sticking to agreements. But this leads to a "contagion" of the conformists with this "unreliable" norm. The conformists exert only the incentive compatible effort instead of the higher performance commitment $\hat{e}$. Hence, it can indeed be the case as Kreps has informally pointed out that the introduction of high powered incentives might lead to a muddied relationship between employer and employee in the sense that the employee "is sent signals that the relationship is a market exchange and reacts accordingly taking fuller advantages of opportunities presented to him" (Kreps (1997), p. 363).

The principal will nonetheless choose the incentive contract in this equilibrium

[20] The figure is plotted for the same parameter values as figure 2 with the only difference that the fraction of steadfast agents η is 0.87 instead of 0.5.

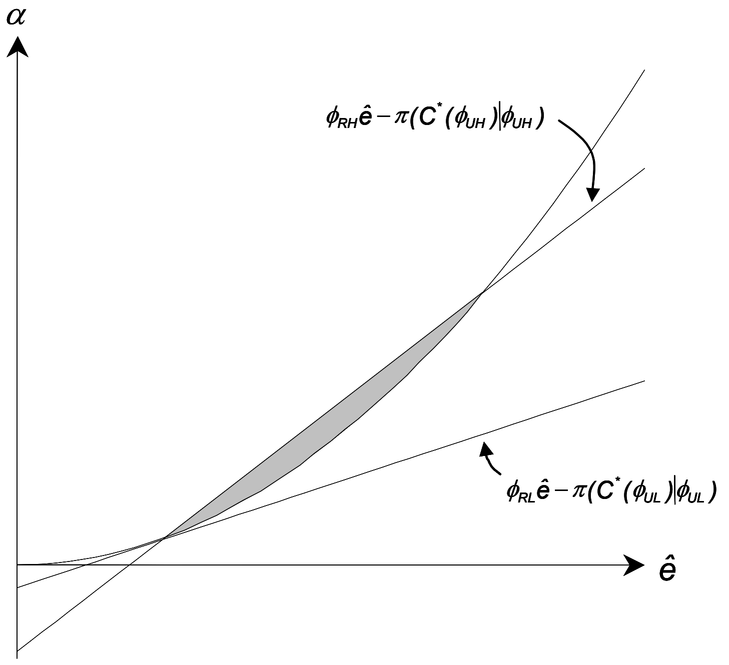

Figure 3: No fixed wage pooling equilibria

as here the choice of a fixed wage contract is not attractive: Although this would lead the conformists to behave reliably she is still better off *in expected terms* when choosing an incentive contract when her signal indicates a high fraction of unreliable agents among the steadfast.

It is important to note that situations will arise in which the principal due to a misleading signal chooses an incentive contract but would have been better off with a fixed wage contract. Consider for instance the extreme case where all steadfast agents are either reliable or unreliable and the principal's signal gives an indication which of the two cases holds but this signal is wrong with a given probability smaller than a half. Furthermore, assume that the prior belief is such that the conformists are initially reliable. When now the principal mistakenly receives the low signal although in truth the steadfast agents are all reliable, she proposes an incentive scheme. She will then be worse off with this decision for two reasons:[21] On the one hand, she provides costly incentives for the steadfast agents although they would have worked hard with mere fixed wages. But in addition, the conformists motivation to stick to agreements is crowded

[21] Still, of course this decision will be optimal given her information.

out. Hence, if the management of a firm underestimated the work ethics of its employees and therefore introduces an incentive scheme it may indeed turn out that overall performance is actually reduced.

5 Conclusion

The sociologist Max Weber introduced a distinction among *instrumental rationality* ("Zweckrationalität") and *value rationality* ("Wertrationalität") when describing the determinants of social behavior. Individuals behave *instrumentally rational* when they choose an action to achieve some goal or end. But their behavior may also be determined by *value rationality* that is by „the conscious belief in the unconditional [..] value of a certain form of behavior purely for its own sake and independent of its success"[22]. Contract theoretical models typically assume purely instrumentally rational behavior: Agents do not stick to agreements specifying unverifiable actions as their preferences are unaffected by the agreement itself but only by the consequences of their actions.

However, some people attach an inherent value to the fulfillment of a promise and are therefore driven by value rationality in this respect. As we have seen, such behavior can be analyzed with standard utility theory simply by imposing that such an agent incurs a huge utility loss when breaking promises. It has turned out that optimal incentive contracts are lower powered than those analyzed in standard hidden action models as long as there is a positive probability that an agent is reliable.

But the value rationality of people might be influenced by *social norms*, i.e. their beliefs on the value others attach to certain actions, namely in our context to the question on whether to stick to unverifiable agreements. As we have seen, incentive contracts then affect the preferences of people as the introduction of a costly incentive scheme signals that not sticking to agreements is a widespread and therefore acceptable behavior. This leads to motivation crowding-out among those agents whose behavior is indeed affected by social norms. Instead of a higher agreed (and "potentially value rational") effort level they choose only a lower, incentive compatible ("instrumentally rational") level. If the principal for instance mistakenly thinks that the fraction of agents having firm preferences

[22] Compare Weber (1956), p. 17, translation by the author.

for the fulfillment of agreements is small, a situation may arise in which she is actually worse off when proposing an incentive scheme as compared to a pure fixed wage.

On a more general level, our theory is therefore in line with experimental evidence found by Gneezy and Rustichini (2000b) who compared the effect of different payment schemes on the performance of children when collecting for a charity. The amount collected by those groups who received a commission rose with the height of the commission. Hence, the incentive effect is present and works well in line with standard contract theory. But the highest amounts were collected by the groups who did not receive any commission at all. It therefore seems that the fact that a commission was paid shifted the reason why the children collected the money. Whereas they might have felt committed to collect as much money as possible without monetary incentives, the introduction of a commission per se may have led them to orientate their actions towards a more "incentive compatible" effort level.

These theoretical considerations have of course some important practical implications. As we have seen, already in our simple model a multiplicity of equilibria arises if the proportion of conformists is sufficiently large. This may give some hints explaining the diversity of successful human resource management policies in practice. On the one hand, there are prosperous firms, as has for instance been stressed by Baron and Kreps (1999)[23], that have established a corporate culture largely based on commitments without strong individual performance incentives. But other firms work with high powered individual incentive schemes and may be successful as well with a "pay-for-performance" culture, in which employees expect direct monetary rewards for higher performance and receive those rewards. Both types of corporate culture can be stable situations and may arise as equilibrium outcomes.

But the mechanisms described indicate a certain danger of "playing around" with incentive schemes. A firm may for instance consider the implementation of an incentive scheme and may therefore want to test the scheme for a trial period and abandon it if it does not bring about the expected benefits. But the danger of a possibly transitory introduction of monetary incentives is that the system of social norms may be shifted away from a commitment based culture to a culture

[23] See for instance the discussion and examples in Chapters 3 and 11.

where agents expect direct rewards. Such a shift in culture may be persistent as the fraction of agents sticking to their performance commitment may be reduced and may stay at the lower level even after the incentive scheme is abandoned again. Such an effect has indeed been empirically observed again by Gneezy and Rustichini (2000a), where the impact of a fine imposed on parents for coming late when picking up their children from a day-care center had been examined. It was not only the case that the introduction of the fine increased the number of late coming parents, but even after the final cancellation of the fine, the number of late comers remained at a higher level than before the introduction.

Furthermore, our model indicates that the introduction of an incentive system can have a very diverse impact on individual behavior. It will unambiguously raise the effort of those employees who do not feel obliged to stick to performance commitments from the beginning. However, it may well lower the motivation of employees who previously felt bound to those commitments. The latter employees may learn that such a commitment is mere "cheap talk" and therefore orientate their efforts towards a lower level which is optimal from the perspective of their material self-interest.

6 Appendix

Proof of Proposition 2:
The principal maximizes

$$\begin{array}{ll} \max\limits_{\beta,\hat{e},e} & (1-\beta)\left[\phi\hat{e}+(1-\phi)\,e\right]+\beta\hat{e}-\frac{c}{2}\hat{e}^2-\frac{1}{2}r\sigma^2\beta^2-k \\ \text{s.t.} & \beta-ce=0. \end{array}$$

Substituting $e=\beta/c$ yields

$$\max_{\beta,\hat{e}}(1-\beta)\left[\phi\hat{e}+(1-\phi)\frac{\beta}{c}\right]+\beta\hat{e}-\frac{c}{2}\hat{e}^2-\frac{1}{2}r\sigma^2\beta^2-k. \tag{10}$$

The optimum is characterized by the following two first order conditions :

$$\begin{aligned} (1-\beta)\,\phi+\beta-c\hat{e} &= 0, && (11) \\ -\phi\hat{e}+(1-\phi)\frac{1}{c}(1-2\beta)+\hat{e}-r\sigma^2\beta &= 0. && (12) \end{aligned}$$

For a given β the performance commitment $\hat{e}$ is obtained from equation (11):

$$\hat{e} = \frac{\beta + (1-\beta)\,\phi}{c} \tag{13}$$

Note that this yields that $\hat{e} > e = \beta/c$. By inserting this expression for $\hat{e}$ into equation (12) and solving for β we then obtain

$$\beta = \frac{1-\phi^2}{1-\phi^2+cr\sigma^2}. \tag{14}$$

The value of e is simply $e = \beta/c$, that of $\hat{e}$ is directly obtained by inserting (14) into (13):

$$\frac{1}{c}\left(\frac{1-\phi^2}{1-\phi^2+cr\sigma^2} + \left(1 - \frac{1-\phi^2}{1-\phi^2+cr\sigma^2}\right)\phi\right) = \frac{1-\phi^2+\phi cr\sigma^2}{c\left(1-\phi^2+cr\sigma^2\right)}.$$

The fixed wage α can then be computed by solving the binding participation constraint (1) for α:

$$\begin{aligned}\alpha &= -\beta\hat{e} + c\left(\hat{e}\right) + \frac{1}{2}r\sigma^2\beta^2 \\ &= \frac{\left(cr\sigma^2 - 1 + \phi^2\right)\left(cr\sigma^2\phi^2 + 1 - \phi^2\right)}{2c\left(1-\phi^2+cr\sigma^2\right)^2}\end{aligned}$$

■

Proof of Proposition 3:

With pure fixed wages the principal's profit is simply

$$\pi\left(C_F^*\left(\phi\right)|\phi\right) = \phi\frac{\phi}{c} - \frac{\phi^2}{2c} = \frac{\phi^2}{2c}.$$

When she instead chooses the optimal incentive contract, we can compute profits by inserting the expression for $\hat{e}$ as a function of β given by (13) into principal's objective function (10) and simplifying, we get the total profit as a function of β

$$\frac{1}{2c} - \frac{\left(1-\phi^2\right)\left(1-\beta\right)^2}{2c} - \frac{1}{2}r\sigma^2\beta^2 - k,$$

and finally by inserting the optimal value of β we obtain

$$\begin{aligned}&\frac{1}{2c}-\frac{\left(1-\phi^2\right)\left(cr\sigma^2\right)^2}{2c\left(1-\phi^2+cr\sigma^2\right)^2}-\frac{1}{2}r\sigma^2\frac{\left(1-\phi^2\right)^2}{\left(1-\phi^2+cr\sigma^2\right)^2}-k\\&=\frac{1}{2c}-\frac{\left(1-\phi^2\right)\left(cr\sigma^2\right)^2+\left(1-\phi^2\right)^2cr\sigma^2}{2c\left(1-\phi^2+cr\sigma^2\right)^2}-k,\end{aligned}$$

from where we get that

$$\pi\left(C_V^*\left(\phi\right)|\phi\right)=\frac{1}{2c}-\frac{\left(1-\phi^2\right)r\sigma^2}{2\left(1-\phi^2+cr\sigma^2\right)}-k.$$

If $\phi<1$ this can be rearranged to

$$\frac{1}{2c}-\frac{cr\sigma^2}{2c\left(1+\frac{1}{1-\phi^2}cr\sigma^2\right)}-k$$

from where it can be directly seen that $\pi\left(C_V^*\left(\phi\right)|\phi\right)$ is increasing in ϕ. The principal will choose an incentive scheme if $\pi\left(C_V^*\left(\phi\right)|\phi\right)>\pi\left(C_F^*\left(\phi\right)|\phi\right)$ or

$$\frac{1-\phi^2+\phi^2cr\sigma^2}{2c\left(1-\phi^2+cr\sigma^2\right)}-k>\frac{\phi^2}{2c}\Leftrightarrow\left(\phi^2-(1-ck)\right)^2>c^2k\left(2r\sigma^2+k\right)$$

which is the case if either

$$\begin{aligned}\phi^2&<1-kc-\sqrt{\left(2r\sigma^2+k\right)kc^2}\text{ or}\\\phi^2&>1-kc+\sqrt{\left(2r\sigma^2+k\right)kc^2}.\end{aligned}$$

As $\sqrt{\left(2r\sigma^2+k\right)kc^2}>kc$ the second term is larger than one. Hence, the variable wage will be chosen if

$$\phi^2<1-kc-\sqrt{\left(2r\sigma^2+k\right)kc^2}.$$

If the right hand side is positive the cut-off $\hat{\phi}$ is given by its root, otherwise $\hat{\phi}=0$ and an incentive scheme will never be optimal in that case. ■

Proof of Proposition 4:
(i) If $E[\tau] < \bar{\tau}$ all conformists are unreliable if they receive no further information. Suppose a pooling equilibrium exists with a contract C_P. It is easiest to sustain such an equilibrium if the agent believes that the signal is bad whenever the principal deviates from this equilibrium. In that case, when receiving the high signal the principal optimally chooses $C^*(\phi_{UH})$ where $\phi_{UH} = \eta\tau_H$. Hence, we must have that $C_P = C^*(\phi_{UH})$. But when receiving the low signal, she is better off by choosing the contract $C^*(\phi_{UL})$. We know from Section 3 that for different values of ϕ, different contracts will be optimal. Hence, $C^*(\phi_{UH}) \neq C^*(\phi_{UL})$ and we cannot have a pooling equilibrium with an unreliable prior.
(ii) Take any fixed wage pooling contract $(\hat{e}_P, \alpha_P, 0)$. Again, such an equilibrium is easiest to sustain if the agent believes that the signal is bad (and hence, conformists are unreliable) whenever the principal deviates from this contract. To see this, just note that we have to consider only deviations that are optimal for given beliefs and therefore given probability ϕ of contracting with a reliable agent. But from Proposition 3 we know that the principal's profit from an optimal contract is always increasing in the fraction of reliable agents.

After $s = H$ the most attractive deviation is to choose $C^*(\phi_{UH})$. Hence, we must have that

$$\phi_{RH}\hat{e}_P - \alpha_P \geq \pi\left(C^*(\phi_{UH})\,|\phi_{UH}\right). \tag{15}$$

After observing $s = L$ the most attractive deviation is to offer $C^*(\phi_{UL})$ yielding the necessary condition

$$\phi_{RL}\hat{e}_P - \alpha_P \geq \pi\left(C^*(\phi_{UL})\,|\phi_{UL}\right). \tag{16}$$

Furthermore, for $(\hat{e}_P, \alpha_P, 0)$ to be chosen in equilibrium, it must satisfy the participation constraint $\alpha_P \geq \frac{c}{2}\hat{e}_P^2$.
(iii) We still have to check, whether this set of possible fixed wage pooling contracts is non-empty. We will first show, that this is not the case if η is large. A necessary (but not sufficient) condition for a pooling equilibrium to exist is that the boundary defined by (16) intersects the cost function: Otherwise, contracts satisfying the condition will never fulfill the participation constraint as the boundary is always below the cost function. If an intersection exists, there must be some value of $\hat{e}$ such that $\frac{c}{2}\hat{e}^2 = \phi_{RL}\hat{e} - \pi\left(C^*(\phi_{UL})\,|\phi_{UL}\right)$, which is equivalent

to

$$\left(\hat{e}-\frac{\phi_{RL}}{c}\right)^2=\left(\frac{\phi_{RL}}{c}\right)^2-\frac{2}{c}\pi\left(C^*\left(\phi_{UL}\right)|\phi_{UL}\right).$$

Hence, the intersection exists, whenever

$$\frac{\phi_{RL}^2}{2c}\geq\pi\left(C^*\left(\phi_{UL}\right)|\phi_{UL}\right).$$

If all agents are steadfast ($\eta=1$) then $\frac{\phi_{RL}^2}{2c}=\frac{\phi_{UL}^2}{2c}<\pi\left(C^*\left(\phi_{UL}\right)|\phi_{UL}\right)$ by Assumption 2. Hence, there can be no pooling equilibrium if $\eta=1$. On the other hand, when all are conformists ($\eta=0$) we have that $\phi_{RH}=\phi_{RL}=1$. The contract $C^*(1)$ attains the first-best profit and satisfies both conditions (15) and (16). Hence, a pooling equilibrium always exists if $\eta=0$.[24] The proof is now completed by showing that if there is a fixed wage pooling contract satisfying (15), (16) and the participation constraint for some value of η, there will be also such a pooling contract for all $\eta'<\eta$. To see that take some vector $(\alpha_P,\hat{e}_P)$ satisfying the conditions for some η. For smaller values of η the left hand side of both conditions (15), (16) are larger as ϕ_{RH} and ϕ_{RL} are both decreasing in η. But the right hand sides are smaller as $\pi\left(C^*\left(\phi\right)|\phi\right)$ is increasing in ϕ by Proposition 3 and ϕ_{UH} and ϕ_{UL} are both increasing in η. Hence, the pooling set is larger, the smaller η and a cut-off value for η must exist such that the set is non-empty whenever η is smaller than this cut-off. ■

Proof of Proposition 5:

(i) In a separating equilibrium the agent is informed about the principal's signal. Hence, after learning that the principal received a low signal, a conformist will always be unreliable and the overall probability of contracting with a reliable agent is $\phi_{UL}=\eta\tau_L$. Any contract choice is then dominated by $C^*\left(\phi_{UL}\right)$, which is the full information optimal contract for this case. By Assumption 2 and Proposition 3 this will be an incentive contract.

(ii) In a separating equilibrium the agent infers that the principal received the high signal when she offers C_H. Hence, a conformist will always be reliable in

[24] Note that by the continuity of $\pi\left(C^*\left(\phi_{UL}\right)|\phi_{UL}\right)$ in η pooling equilibria do not exist in the neighbourhood of $\eta=1$. Similarly, for $\eta=0$ the inequalities (15) and (16) are strict. As all functions are continuous, pooling equilibria always exist in the neighbourhood of $\eta=0$.

that case. If the principal has received the high signal she therefore expects a profit of

$$\phi_{RH}\hat{e}_H - \alpha_H$$

when offering the fixed wage contract C_H. Again, if any separating equilibrium can be sustained, then one can be sustained in which the agent believes that the principal received a low signal after any deviation from the equilibrium path. The best deviation for the principal when $s = H$ is to choose the optimal contract for this case, which is $C^*(\phi_{UH})$. Hence, the principal does not want to deviate after $s = H$ if condition (7) holds.

After observing the low signal the principal expects an equilibrium profit of $\pi\left(C^*(\phi_{UL})\,|\phi_{UL}\right)$ as all conformists are unreliable in that case. Note that the only tempting deviation is to imitate the high signal by choosing C_H: Any other deviation will lead the conformists to become unreliable and C_L is already the optimal contract for that case. But with deviation to C_H motivation crowding-out will be prevented and the conformists will become reliable. In that case, the principal's expected profits are $\phi_{RL}\hat{e}_H - \alpha_H$. She will not have an incentive to deviate to C_H if condition (8) holds. Condition (9) is again the participation constraint assuring that the reliable agent's best response to a contract offer C_H is to accept.

(iii) Finally, we have to show that this set of separating equilibria is always non-empty. It will be shown that there is always at least one vector (α, e) satisfying all three conditions. We consider the upper intersection of the upper bound with the cost function.[25]

$$\frac{c}{2}e^2 = \phi_{RH}e - \pi^*_{UH} \Leftrightarrow$$
$$\left(e - \frac{\phi_{RH}}{c}\right)^2 = \left(\frac{\phi_{RH}}{c}\right)^2 - \frac{2}{c}\pi^*_{UH}$$

The upper intersection is therefore given by

$$e = \frac{\phi_{RH}}{c} + \sqrt{\left(\frac{\phi_{RH}}{c}\right)^2 - \frac{2}{c}\pi^*_{UH}}.$$

[25] For ease of notation let $\pi^*_{UH} = \pi\left(C^*(\phi_{UH})\,|\phi_{UH}\right)$ and $\pi^*_{UL} = \pi\left(C^*(\phi_{UL})\,|\phi_{UL}\right)$.

Note that the intersection always exists as $\left(\frac{\phi_{RH}}{c}\right)^2 - \frac{2}{c}\pi^*_{UH} > 0 \Leftrightarrow \pi^*_{RH} = \frac{\phi^2_{RH}}{2c} > \pi^*_{UH}$. We have to check, whether this contract is above the lower bound, which is the case if it is to the right of the intersection of both incentive constraints. The intersection of both constraints is given by

$$\phi_{RH}\hat{e} - \pi^*_{UH} = \phi_{RL}\hat{e} - \pi^*_{UL} \Leftrightarrow \hat{e} = \frac{\pi^*_{UH} - \pi^*_{UL}}{\phi_{RH} - \phi_{RL}}.$$

Hence, it has to be verified that

$$\begin{aligned}
&\frac{\phi_{RH}}{c} + \sqrt{\left(\frac{\phi_{RH}}{c}\right)^2 - \frac{2}{c}\pi^*_{UH}} \geq \frac{\pi^*_{UH} - \pi^*_{UL}}{\phi_{RH} - \phi_{RL}} \Leftrightarrow \\
&(\phi_{RH} - \phi_{RL})\frac{\phi_{RH}}{c} + (\phi_{RH} - \phi_{RL})\sqrt{\left(\frac{\phi_{RH}}{c}\right)^2 - \frac{2}{c}\pi^*_{UH}} \geq \pi^*_{UH} - \pi^*_{UL}.
\end{aligned}$$

It suffices to show that

$$(\phi_{RH} - \phi_{RL})\,\phi_{RH} \geq (\pi^*_{UH} - \pi^*_{UL})\,c. \tag{17}$$

Note that $C^*(\phi_{UH})$ can either be a fixed wage or an incentive contract.
a) When $C^*(\phi_{UH})$ is a fixed wage contract, we know that $\pi^*_{UH} = \frac{\phi^2_{UH}}{2c}$. As $\pi^*_{UL} > \frac{\phi^2_{UL}}{2c}$ it suffices to show that

$$\begin{aligned}
&(\phi_{RH} - \phi_{RL})\,\phi_{RH} = \eta\,(\tau_H - \tau_L)\,(1 - \eta + \eta\tau_H) \geq \frac{\eta^2\tau^2_H}{2} - \frac{\eta^2\tau^2_L}{2} \\
&\Leftrightarrow 2\,(1 - \eta + \eta\tau_H) \geq \eta\,(\tau_H + \tau_L) \\
&\Leftrightarrow 2 - 2\eta + \eta\,(\tau_H - \tau_L) \geq 0
\end{aligned}$$

which is always the case.
b) When $C^*(\phi_{UH})$ is an incentive contract we consider the right hand side of

inequality (17) by using the principal's profit as given by Proposition 3:

$$\begin{aligned}
\pi_{UH}^{*}-\pi_{UL}^{*} &= \frac{\left(1-\eta^2\tau_L^2\right)r\sigma^2}{2\left(1-\eta^2\tau_L^2+cr\sigma^2\right)}-\frac{\left(1-\eta^2\tau_H^2\right)r\sigma^2}{2\left(1-\eta^2\tau_H^2+cr\sigma^2\right)} \\
&= \frac{\left(1-\eta^2\tau_L^2\right)\left(1-\eta^2\tau_H^2+cr\sigma^2\right)-\left(1-\eta^2\tau_L^2+cr\sigma^2\right)\left(1-\eta^2\tau_H^2\right)}{2\left(1-\eta^2\tau_L^2+cr\sigma^2\right)\left(1-\eta^2\tau_H^2+cr\sigma^2\right)}r\sigma^2 \\
&= \frac{\left(\tau_H^2-\tau_L^2\right)\eta^2 c\sigma^4}{2\left(\frac{1-\eta^2\tau_L^2}{r}+c\sigma^2\right)\left(\frac{1-\eta^2\tau_H^2}{r}+c\sigma^2\right)}.
\end{aligned}$$

Note that this expression is strictly increasing in r. The more risk averse the agent, the larger is the impact of the difference in the probability of contracting with a reliable agent on profits. Hence,

$$\lim_{r\to\infty}\left(\pi_{UH}^{*}-\pi_{UL}^{*}\right)c=\frac{\left(\tau_H^2-\tau_L^2\right)\eta^2}{2}$$

constitutes an upper bound on the right hand side of inequality (17). From part (a) we know that therefore (17) is always met. ■

References

Akerlof, G. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics 97*, 543–69.

Alger, I. and R. Renault (2000). Screening among agents with heterogeneous ethics. *Boston College, Department of Economics, Working Paper 489*.

Baron, J. N. and D. M. Kreps (1999). *Strategic Human Resources: Frameworks for General Managers.* New York: John Wiley & Sons.

Bem, D. J. (1967). Self-perception: The dependent variable of human performance. *Organizational Behavior and Human Performance 2*, 105–121.

Bénabou, R. and J. Tirole (2002). Intrinsic and extrinsic motivation. Mimeo IDEI Toulouse.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity and social history. *Games and Economic Behavior 10*, 122–142.

Bolton, G. and A. Ockenfels (2000). ERC - a theory of equity, reciprocity and competition. *American Economic Review 90*, 166–193.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*, 817–869.

Datar, S. M., S. Kulp, and R. Lambert (2001). Balancing performance measures. *Journal of Accounting Research 39*, 75–92.

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

Deci, E. L. and J. Porac (1978). Cognitive evaluation theory and the study of human motivation. In M. R. Lepper and D. Greene (Eds.), *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*, pp. 149–175. Hillsdale, NJ: Lawrence Erlbaum.

Dufwenberg, M. and G. Kirchsteiger (1999). A theory of sequential reciprocity. *Mimeo, CentER for Economic Research, Tilburg.*.

Falk, A. and U. Fischbacher (1999). A theory of reciprocity. *Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6*.

Fehr, E. and S. Gächter (2002). Do incentive contracts crowd out voluntary cooperation? Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 34.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997). Reciprocity as a contract enforcement device - experimental evidence. *Econometrica 64*, 833–860.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does fairness prevent market clearing? an experimental investigation. *Quarterly Journal of Economics 108*, 437–460.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*, 817–868.

Fehr, E. and K. M. Schmidt (2002). Theories of fairness and reciprocity - evidence and economic applications. In M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.), *Advances in Economics and Econometrics - 8th World Congress.*

Feltham, G. and J. Xie (1994). Performance measure congruity and diversity in multi-task Principal/Agent relations. *The Accounting Review 69*, 429–453.

Frey, B. S. (1997). *Not Just For the Money. An Economic Theory of Personal Motivation.* Cheltenham: Edward Elgar.

Frey, B. S. and L. Götte (1999). Does pay motivate volunteers? *Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 9*.

Frey, B. S. and R. Jegen (2001). Motivation crowding theory: A survey of empirical evidence. *Journal of Economic Surveys 15 (5)*, 589–611.

Frey, B. S. and F. Oberholzer-Gee (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review 87*, 746–755.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological games and sequential rationality. *Games and Economic Behavior 1*, 60–79.

Gibbons, R. S. (1997). Incentives and careers in organizations. In D. M. Kreps and K. Wallis (Eds.), *Advances in Economic Theory and Econometric: 7th World Congress of the Econometric Society*, Cambridge. Cambridge University Press.

Gibbons, R. S. and K. J. Murphy (1992). Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence. *Journal of Political Economy 100*, 468–505.

Gneezy, U. and A. Rustichini (2000a). A fine is a price. *Journal of Legal Studies 29*, 1–18.

Gneezy, U. and A. Rustichini (2000b). Pay enough or don't pay at all. *Quarterly Journal of Economics*, 791–810.

Holmström, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization 7*, 24–52.

Holmström, B. and P. Milgrom (1994). The firm as an incentive system. *American Economic Review 84*, 972–991.

Holmström, B. and P. Milgrom (1990). Regulating trade among agents. *Journal of Institutional and Theoretical Economics 146*, 85–105.

Irlenbusch, B. and D. Sliwka (2003). Incentives, decision frames and motivation crowding out - an experimental investigation. *Mimeo, Universität Bonn.*

Koford, K. and M. Penno (1992). Accounting, principal-agent theory, and self-interested behavior. In N. E. Bowie and R. E. Freeman (Eds.), *Ethics and Agency Theory*, New York, pp. 127–142. Oxford University Press.

Kreps, D. M. (1997). Intrinsic motivation and extrinsic incentives. *American Economic Review 87 (2)*, 359–64.

Kunz, A. and D. Pfaff (2002). Agency theory, performance evaluation, and the hypothetical construct of intrinsic motivation. *Accounting, Economics and Society 27*, 275–295.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review 90 (5)*, 1346–62.

Lepper, M. R. and D. Greene (1978). Overjustification research and beyond: Towards a means-ends analysis of intrinsic and extrinsic motivation. In M. R. Lepper and D. Greene (Eds.), *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*, pp. 109–148. Hillsdale, NJ: Lawrence Erlbaum.

Parent, D. (2001). Incentive pay in the united states: Its determinants and its effects. *University of Montreal, Cirano Working Paper 2001s-04.*

Parent, D. and W. B. MacLeod (1999). Job characteristics and the form of compensation. *Research in Labor Economics 18.*

Prendergast, C. J. (1999). The provision of incentives in firms. *Journal of Economic Literature 37*, 7–63.

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

Weber, M. (1956). *Wirtschaft und Gesellschaft.* Tübingen: J. C. B. Mohr.