ON BUILDING A CONFLICT CULTURE IN ORGANIZATIONS

W. BENTLEY MACLEOD, VICTORIA VALLE LARA, AND CHRISTIAN ZEHNDER

Although conflicts are typically associated with negative emotions and waste of resources, organizations may still benefit from a corporate culture that tolerates or even encourages conflict by workers. The reason is that coordinated conflicts can help to enforce informal contracts and foster cooperation. In this paper we report results of a series of laboratory experiments designed to explore whether and under what conditions an efficiency-enhancing conflict culture can emerge. Using a principal-worker setup with subjective performance evaluation, we show that establishing a functional conflict culture is a delicate matter. If conflicts are encouraged in a careless, hands-off manner, the destructive side of conflicts is likely to dominate. To be successful a conflict culture requires a careful management of fairness norms. In our experiment we find that conflicts have positive net effects on efficiency only if an explicit code of conduct is established and conflicts are institutionalized through a grievance process. Thus, providing workers with more power may be a necessary but not sufficient condition for improving productivity when performance evaluations are subjective.

Date: 2019-10-27.

"A good manager doesn't try to eliminate conflict; he tries to keep it from wasting the energies of his people. If you're the boss and your people fight you openly when they think that you are wrong–that's healthy."

Robert Townsend (2007)

1. Introduction

A corporate culture that tolerates or even encourages conflicts in case of perceived injustice entails both opportunities and dangers. On one hand, a healthy conflict culture may establish credible threats that motivate otherwise non-complying parties to respect social norms and to contribute to the greater good. On the other hand, a conflict-friendly culture may also create escalations, negative emotions and waste of resources. In this paper we study the role of conflicts for employment relationships in which compensation depends on subjective performance evaluation. We report evidence from a series of laboratory experiments illustrating that the formation of an efficiency-enhancing conflict culture is a delicate matter. If conflicts are encouraged in a careless, hands-off manner, the destructive side of conflicts is likely to dominate. A functioning conflict culture requires a careful management of the informal contracts between individuals in the organization.[1] In our experiment we find that conflicts have positive net effects only if an explicit code of conduct is established and conflicts are institutionalized through a grievance process.

There is ample evidence illustrating that perceived unfairness is an important trigger of conflicts. Many people are willing to engage in costly retaliation in response to observed violations of implicit agreements that are anchored by norms. Famous examples of such behavior include rejections of small offers in ultimatum games (Guth et al., 1982; Kahneman et al., 1986), punishment of free-riders in public goods experiments (Gächter and Fehr, 2000; Ostrom et al., 1992), or third-party interventions in distribution games (Leibbrandt and López-Pérez, 2012; Gächter and Fehr, 2000). Although such conflicts are destructive and inefficient in the short term (Dreber et al., 2008; Egas and Riedl, 2008), they can be ultimately efficiency-enhancing because—once established—the threat of conflicts may induce norm breakers to comply and cooperate (Gächter et al., 2008).

Existing studies reporting a strong disciplining and cooperation-enhancing effect of conflicts have predominantly focused on deterministic set-ups with symmetric information (see, Chaudhuri, 2011, for a survey).[2] Those findings may not generalize to the employment context where relationships are typically characterized by hidden actions, stochastic input-output links, and noisy information. The presence of probabilistic outcomes and information asymmetries may considerably complicate the role of conflicts, because there may be disagreement about what constitute fair behavior and violations of norms may become harder to detect. Our study presents a novel experimental design that sheds new light on the empirical relevance of those important issues.

The game underlying our experiments builds on a two-person, principal-worker setting with subjective performance evaluation. The Worker's effort is private information and creates a stochastic output that neither the Principal nor to the Worker can observe. Instead, the Principal and the Worker each get a private, subjective signal that gives imperfect information about the output created by the Worker. These signals are noisy and imperfectly correlated so that the Principal and the Worker sometimes have contradicting information about output. Given that signals are private information, it is not possible to contract on them.

[1] Jensen and Meckling (1976) famously observed that a firm is nothing more than a "nexus of contracts".

[2] Grechenig et al. (2010) and Ambrus and Greiner (2012) show that peer-punishment technologies are less beneficial in public good games in which players have imperfect information about contributions of others.

After observing the subjective signal, principals can therefore simply decide whether or not to pay a bonus to the Worker. To study the impact of conflicts on efficiency, we compare two different treatments. In the no conflict condition, the game ends after the Worker observes the bonus choice of the Principal. In the conflict condition, the Worker can start a conflict with the Principal after observing the bonus choice. Conflicts are implemented in the form of costly sanctions. This reflects the view that the immediate consequence of conflict is typically negative for all involved parties, but mostly so for the party under attack.

So what should the players optimally do in our setup? Efficiency requires that workers exert high effort. Principals, in turn, can use bonus payment to reward workers and establish a fair allocation of payoffs. Unfortunately, such a simple reciprocal exchange is unlikely to occur. It is a well-established fact in the social preference literature that there is large heterogeneity in the degree to which people care about social motives such as fairness. In particular, whereas some people go out of their way and make considerable sacrifices to enforce social norms, there is also a considerable part of the population whose behavior is best described as approximately selfish (see, e.g., Cooper and Kagel, 2016, for a recent survey of the literature). In the absence of conflicts, the presence of selfish principals who have no motive to pay bonuses undermines a well-functioning exchange of gifts. Intuitively, conflicts can help to mitigate this problem: if sufficiently many workers sanction principals who violate fairness norms, even selfish principals may have an incentive to reward workers appropriately (Fehr and Fischbacher, 2004, 2003).

A complicating issue is that it is not obvious what a fair compensation entails in our setup. Previous work on bargaining games shows that perceived fairness is both context specific and shaped by strategic considerations (Prasnikar and Roth, 1992; Binmore et al., 1993) and these effects are even more pronounced under uncertainty (Kagel et al., 1996). In our environment uncertainty manifests itself in the stochastic link between the Worker's input and output and may imply that multiple fairness norms co-exist (see also Akerlof, 1980).[3] One view is that fairness requires that agents are compensated for exerting high effort (the *pay-for-input norm*). Alternatively, one might argue that agents should be rewarded if output turns out to be high (the *pay-for-output norm*).[4] For a functioning conflict culture it is crucial that the Worker and the Principal agree on the fairness norm on which their implicit agreement builds. If trading partners (unknowingly) disagree about the definition of norm-abiding behavior, misunderstandings will be interpreted as breach of agreement and may lead to wasteful conflicts. Moreover, as it turns out, it not only matters that parties agree on some equilibrium norm, but the efficiency of their implicit agreement also depends on the particular norm they coordinate on.

The *pay-for-input (PI) norm* prescribes that the Worker provides effort and the Principal reciprocates with bonus pay. Workers who enforce the *PI norm* ignore their subjective output signal and view any non-payment of the bonus as breach of agreement. If there are sufficiently many norm enforcers, the threat of conflict will induce principals to pay the bonus irrespective of their own performance signal. The advantage of implicit agreements that build on the *PI norm* is that cooperation does not require any conflicts on the equilibrium path. Principals are disciplined by the threat of conflict alone. The downside, on the other hand, is that the incentive structure established by the *PI norm* fails to motivate selfish agents to exert high effort. In fact, unconditional bonus payments invite selfish agents to invade the system and to free-ride on

[3]MacLeod and Malcomson (1998) discuss the existence of multiple equilibrium norms in labor markets within the context of self-enforcing relational contracts.

[4]The labels we use for fairness norms refer to Lazear's (1986) categorization of compensations schemes. However, whereas Lazear's work aims at identifying determinants of different compensation forms, our interest is in exploring the efficiency consequences of norm enforcement in employment relationships.

the norm enforcement of others. If the population fraction of selfish agents is sufficiently large, the *PI norm* will therefore not lead to a large increase in performance.

The upside of the *pay-for-output (PO) norm* is that it deters the shirking of self-interested agents by connecting performance to pay. When there is subjective evaluation, as in our case, neither the Principal's nor the Worker's choices can be contingent on objective output. However, the Principal pays a bonus if and only if she observes a high subjective performance signal and the Worker feels entitled to a bonus payment if and only if his own subjective signal suggests that output is high. If a sufficiently large population of agents enforce the pay-for-output norm, principals have an incentive to make signal-contingent bonus payments and all agents (including those who are selfish) have an incentive to provide high effort. However, since the Worker's signal of performance is imperfectly correlated with the Principal's signal, the two parties may receive conflicting signals. In particular, the Principal might observe a low signal, and not pay the bonus, while the Agent observes a high signal and expects to get a bonus. In this case, the *PO norm* implies that the Worker believes that the Principal is in breach of agreement, which gives him the right to punish the Principal. Thus, there will be conflict on the equilibrium path. However, as long as the probability of contradicting signals is not too high, the damage created by occasional conflicts is largely dominated by the efficiency gains from full effort provision.

Much of the literature on behavioral contract theory focuses upon how social preferences shape behavior and contract design (Fehr and Schmidt, 1999; Fehr et al., 2011; Koszegi, 2014). This research takes the population of workers and their preferences as given and tries to adjust contracts and institutions such that performance is high. In this paper, in contrast, we take the view that preferences are partially endogenous in that they may depend on the institutional environment. Specifically, we assume that the details of the context may affect the selection of fairness norms that trading parties use as the basis of their implicit agreements. We explore the coordination on fairness norms within the context of two treatment conditions– conflict and no conflict. Thus the first question we ask is whether or not the potential for conflict can enhance organizational performance. Second, we compare these treatments across varying contexts, while holding the relationship between actions and payoffs fixed. In standard behavioral contract models where preferences are defined exclusively over payoffs and the distribution of payoffs between the two parts the variations in context should have no effect. If they do (and we find that they do), then this is evidence that context (e.g. in the form of "corporate culture") can affect behavior.

Our experiments aim at exploring whether and under what conditions an efficiency-enhancing conflict culture based on the pay-for-output norm can emerge. In our first experiment (henceforth called "baseline experiment") participants spontaneously interact in our subjective evaluation setup without any further intervention or clarification. We compare two treatments in a between-subjects design: in one condition conflicts are absent by design, in the other condition workers can endogenously engage in costly conflicts after having observed the Principal's bonus decision. In both treatments participants play the game for 15 periods with changing partners (random matching protocol). We observe that participants fail to build a functioning conflict culture in the baseline experiment. The conflict treatment increases the frequency of bonus payments and leads to a small, but statistically insignificant increase in effort. However, overall, the damage inflicted by the emergence of conflicts is more important than the small increase in productivity, so that total surplus decreases in the presence of conflicts. The underlying problem is that participants fail to coordinate on the pay-for-output norm. About half of the workers in the conflict treatment induce costly sanctions, but many of those norm enforcers exhibit sanction patterns that are more in line with the

pay-for-input norm rather than the pay-for-output norm. The multiplicity of norms leads to uncoordinated conflicts with only limited motivating effects on workers.

Our baseline experiment teaches us an interesting lesson: a careless, hands-off implementation of a conflict culture can easily backfire. Uncoordinated conflicts are costly and have only a very limited positive impact on motivation. At the same time, one might argue that the lack of any communication makes coordination "unnaturally" complicated. In most real-life settings communication opportunities are available and organizations can encourage conversations on conflict to foster the development of a broadly accepted fairness norms. Our second experiment (subsequently termed "communication experiment") investigates the extent to which communication helps the trading parties to coordinate on an efficiency-enhancing use of conflicts. To this purpose, we added an additional stage at the beginning of every period of the experiment. In this communication stage principals and workers select a message from a pre-defined set of messages and send it to their trading partner. The set of available messages consisted of strategy announcements (principals: bonus payment strategy / workers: effort choice and conflict initiation strategy) and included the strategies underlying the two fairness norms plus some other alternatives.

The results of our second experiment reveal that the communication does not necessarily facilitate the establishment of an efficiency-enhancing conflict culture. As in the baseline experiment, average effort in the conflict treatment is only slightly higher than in the no conflict treatment and overall surplus decreases in the presence of conflicts. All in all, it seems that communication further reinforces rather than mitigates the problems observed in the baseline experiment. In particular, communication leads to an even stronger use of the pay-for-input norm than in the baseline experiment and the overall negative impact of conflicts on efficiency remains.

The communication experiment confirms that self-coordination on an efficiency-enhancing fairness norm seems difficult in our setting. As a next step we study the impact of a direct external appeal to coordinate on the pay-for-output norm. Our third experiment (henceforth called "the agreement experiment") explores the effectiveness of a code of conduct in our set-up. In this experiment we begin each period with a stage in which both the Principal and the Worker are asked to electronically sign a non-binding agreement in which they confirm their intention to follow the pay-for-output norm. Failure to sign the agreement by at least one party implies that trade does not occur and an unattractive outside option is implemented. Surprisingly, this intervention has only very weak effects. Despite the fact that the vast majority of participants sign the agreement, most players do not seem to feel obliged to respect the rules of the code of conduct. As a consequence, the results of the agreement experiment are similar to those of the baseline experiment.

In our fourth and final experiment (subsequently called "the grievance experiment") we extend the agreement experiment by adding one decisive feature: institutionalized conflicts. In this experiment the trading parties sign the same code of conduct as in the agreement experiment. The only difference is that in this experiment the initiation of conflict requires a formal grievance process. If workers intend to engage in a conflict with their principal they need to file a complaint in which they explicitly confirm that the Principal has violated the code of conduct. It is important to emphasize that in our experiment the institutionalization of the conflict is just a re-framing of the same action space as the agreement experiment. In particular, there is no verification process after an worker files a complaint and there is no material consequence to lying. Nevertheless, we observe that the outcomes in the grievance experiment are substantially different from those in our other three experiments. In this experiment, the trading parties follow more frequently the pay-for-output norm and conflicts emerge in a much more coordinated manner. Moreover, the conflict

patterns also induce the principals to make their bonus payments contingent on their subjective signal so that even selfish workers have a monetary incentive to exert high effort. The treatment comparison in this experiment reveals a significant increase in the average effort level and higher total surplus in the conflict treatment relative to the no conflict treatment.

Our results have implications for conflict management in organizations. The management literature recognizes that conflicts are inevitable and many authors recommend tactics and procedures to handle conflicts constructively (Coleman et al., 2014; De Dreu et al., 2008; Tjosvold et al., 2014). Our findings suggest that even conflicts that appear to be purely wasteful may be ultimately efficiency-enhancing if they occur in a coordinated manner. The reason is that the presence of these conflicts establishes an implicit incentive system that increases cooperation. The challenging part for management is the establishment of broadly shared fairness norms on which implicit agreements among individuals within the organization are built. We think that the effective choice of managerial interventions that coordinate people on specific norms is part of what creating a corporate culture means. Our view is closely related to the approach of identity economics which argues that individuals' preferences should no longer be modeled as fixed but rather as as a function of social context (Akerlof and Kranton (2000)). The assumption is that people not only care about outcomes, but are also concerned about the extent to which behavior is compatible with social norms they identify with.

Our work extends the literature on norms in two ways. On the theory side, we introduce the notion of a norm equilibrium. This concept allows us to establish the behavioral stability of a particular fairness norm with respect to invasion from players following another norm (including players who follow the rational, self-interest "norm"). As we point out above, our experiment is simple enough that the literature on the economics of pay identifies three natural norms: pure self-interest, pay-for-input exchange, and pay-for-output exchange. Our parameter values are set such that in the no conflict treatment only the pure self-interest norm is an equilibrium. In the conflict treatment, in contrast, both the pure self-interest norm and the pay-for-output norms are equilibria and the pay-for-output equilibrium is the payoff-dominant one. This result motivates our quest for contextual conditions that favor coordination on the pay-for-output norm. On the empirical side, we demonstrate that a feature of our environment is that outcomes can be represented with a binary tree. If subjects are assumed to choose a norm with error, then the probability distribution of the end nodes of the tree can be represented by a multi-nomial logit. Hence adherence to a norm can be estimated using an off the shelf multi-nomial regression. This approach not only allows us to measure whether a particular norm gets better or worse at describing the data as we move across environments, but it also enables us to use a non-nested likelihood ratio test to formally test which norm does the best job of fitting the data within a given treatment. We see this as a contribution to the development of empirical methods for identifying and measuring social norms and cultures.

The remainder of our paper is organized as follows. The next section introduces the framework, including the payoffs for each party and the definition of a norm equilibrium. Section 3 presents the experimental results by environment. We measure the effect of allowing conflict upon effort and welfare in each of the contexts that we implemented in the laboratory. In Section 4 we introduce a method to measure behavior and the prevalent social norm in the experiment. We provide a formal test of which norm best fits the data.

2. Framework

As the basis for our analysis we use a simplified version of the subjective evaluation model introduced in MacLeod (Section III). In particular, we consider the following contracting problem: The Principal hires a Worker and offers a contract (w, B), where w is a contractible base wage and B is a non-enforceable bonus that the Principal can pay at her discretion. The Worker can choose an effort $e \in \{0, 1\}$ at cost $c(e) = c \times e$, where $c > 0$. The Worker's effort induces a stochastic return $r \in \{r_H, r_L\}$, where $r_L \geq 0$, $\delta \equiv r_H - r_L > 0$, $Prob(r = r_H|e) = \gamma_e$, and $1 > \gamma_1 > \gamma_0 > 0$. The return is not observable, neither for the Principal, nor for the Worker. However, both the Principal and the Worker receive private and subjective performance signals. The Principal's subjective signal $s_P \in \{0, 1\}$ depends on the return and is determined by the following conditional probabilities:

$$Pr\{s_P = 1|r = r_H\} = p,$$
$$Pr\{s_P = 1|r = r_L\} = 1 - p,$$

where $p \in [0.5, 1)$. The Worker's subjective signal $s_A \in \{0, 1\}$ depends on the Principal's signal and the corresponding conditional probabilities are:

$$Pr\{s_A = 1|s_P = 1\} = q.$$
$$Pr\{s_A = 1|s_P = 0\} = 1 - q,$$

where $q \in [0.5, 1]$.

After having observed her private signal, the Principal decides whether or not she pays the bonus ($b \in \{0, B\}$) to the Worker on top of the contractually agreed upon base wage w. The Worker observes the bonus decision of the Principal and then decides whether or not to initiate a conflict $d \in \{0, 1\}$ that reduces the Principal's payoff by an amount $f(d) = dF$, where $F > 0$. Engaging in a conflict also has a material cost for workers K, but workers may experience a subjective, psychological utility v if they sanction a Principal that they believe has violated a fairness norm. The total cost of initiating a conflict for the Worker is therefore $k(d, v) = dK - v$, where $0 < K \leq F$ and $v \geq 0$.

Thus, the Principal and the Worker realize the following payoffs:

$$\pi_P = r_L + \gamma_e\delta - w - b - f(d),$$
$$\pi_A = w + b - c(e) - k(d, v).$$

The decisive parameters of this game are γ_1, γ_0, p, and q. These parameters govern the stochastic relation between input (effort) and output (return), the degree to which the Principal's signal is distorted, and the subjectiveness of the players' performance evaluation (i.e., the correlation between the Principal's and the worker's signal). If $q = 1$ the signals of the Principal and the Worker are always perfectly aligned and performance evaluation is no longer subjective. If $p = 1$ the Principal has perfect information about the worker's output (thus, the special case $p = q = 1$ corresponds to the standard principal-worker model). Finally, if $\gamma_1 = 1$ and $\gamma_0 = 0$, input and output are perfectly correlated so that the contracting problem becomes trivial and first best effort can be implemented by a simple output-contingent contract.

To study the impact of conflicts on efficiency, we compare two different treatments. In the no conflict condition, the game ends after the Worker observes the bonus choice of the Principal. In the conflict condition, the Worker can start a conflict with the Principal after observing the bonus choice.

Norm Guided Behavior. The economic approach to behavior in such an environment typically relies upon some version of a Nash equilibrium. Each party chooses a strategy that maximizes their own material payoff given correct expectations of the other party's behavior. Decades of research has shown that individuals do respond to material incentives, much as predicted by the theory. However, there is also decades of research showing that individuals respond imperfectly. They make mistakes, do no consider all the options, get upset and process information imperfectly. One of the jobs of management is to help organizations perform better, given the empirical fact that they must work with imperfect humans.

A fundamental reason that individuals do not always follow the optimal choice is that the strategy space they face is huge.[5] Even though our experiment is a simplified version of employment, we nevertheless observe all sorts of behavior. One of the ways this problem is solved in practice is to provide guidance on appropriate behavior. It is natural for humans to aspire to work hard and follow accepted norms. Technically, we can think about strategies as a bundle of behaviors that "makes sense" and fit together naturally. It is not at all obvious what is "natural" in a particular context, but in our experiment we find that we can organize our results around three norms. In this paper, we first ask the question, is there a set of norms, which can be identified as a "culture", that forms an efficient and stable equilibrium? Second, our experiment measures the extent to which each treatment moves behavior to one culture or the other.

As we discussed in the introduction, Lazear (1986) identifies two norms that are natural in a work context. The first we call *pay-for-input*. When the Principal and Worker meet, the Worker agrees to provide effort in exchange for pay. This norm is a version of reciprocity behavior that has been widely documented in behavioral economics[6]. What we know is that if the Worker provides effort with cost X, but the Principal does not pay the bonus, then the Worker feels unfairly treated, and will, if given the chance, reciprocate by punishing the Principal.

Formally, let σ_W^{PI} and σ_P^{PI} be the strategies in the game described above corresponding to *pay-for-input* or PI for the Worker and the Principal. The PI entails the Worker choosing high effort and the Principal making an unconditional payment of a bonus B. If the bonus is not paid, then the Worker, if she is in a conflict treatment, chooses to punish the Principal with $d = 1$.

The problem is that this norm is unstable as the level of effort supplied is not observed and the Principal cannot verify that effort is in fact supplied. One can immediately see the behavioral problem: the Worker has a large incentive to shirk. Classical agency theory predicts that if effort cannot be observed, then the optimal contract should be one in which the Worker is rewarded for observed performance measures. In our experiment the Principal can observe a signal of performance. We dub *pay-for-output*, or simply PO, as the social norm that entails having the Principal pay a bonus, if and only if, she observes a high signal. The twist, relative to agency theory, is that this signal is *subjective* and not observed by the Worker.

Thus, even if the Principal uses her signal to determine the bonus, the Worker cannot verify that the Principal will pay a bonus after observing a high subjective performance signal. This is solved by the fact that the Worker has his own performance signal. In that case, if the Worker observes a low signal, then he knows that he does not deserve a bonus. However, if he observes a high signal and the Principal does not pay a bonus, then he is aggrieved and feels that the Principal has not treated him fairly. As a consequence he chooses to punish the Principal with $d = 1$. We call this combination of behaviors the *pay-for-output* or PO culture. Formally, let σ_W^{PO} and σ_P^{PO} be the strategies in the game described above corresponding to *pay-for-output* or PO for Worker and the Principal.

[5]This is well known since the seminal work of Simon (1982).

[6]See Fehr and Schmidt (1999) and Falk et al. (2003) for example

Finally, we consider the completely "cynical" players. In this experiment it is costly to initiate conflict and, hence, a cynical/self-interested Worker would never punish the Principal for being unfair. Given this, the cynical/self-interested Principal would face no consequences to not paying a bonus and, hence, would never pay. The Worker will not be rewarded for high effort and, hence, will shirk. This is the unique sub-game perfect Nash equilibrium for this game. While some players behave in this way, most do not and, hence, the hypothesis that play can be predicted by the sub-game perfect Nash equilibrium concept is soundly rejected by the data. Let σ_W^{SI} and σ_P^{SI} be the strategies in the game corresponding to *self-interested* choice or "S-I" for short.

In terms of organizing the results we can think in terms of a "norm" equilibrium, motivated in part by Axelrod (1981) and Akerlof (1980). Rather than assume parties can follow *any* strategy, let us suppose that cognitive/behavioral effects imply that parties focus upon one of the three norms. Hence, the strategy space for each player is assumed to be given by:

$$\Sigma_P = \left\{\sigma_P^{PI}, \sigma_P^{PO}, \sigma_P^{SI}\right\},$$
$$\Sigma_W = \left\{\sigma_W^{PI}, \sigma_W^{PO}, \sigma_W^{SI}\right\}.$$

Let $U_P(\sigma_P, \sigma_W)$ and $U_W(\sigma_P, \sigma_W)$ be the corresponding payoffs for the Principal and the Worker. Given this game, the notion of a Nash equilibrium is well defined in "norm space". For this game *self-interest* is always an equilibrium, while *pay-for-input* is never an equilibrium. Finally, *pay-for-output* is an equilibrium, if and only if, we are in the conflict treatment.

In practice we observe lots of noise in the play of individuals, and hence play is not characterized by a single norm. What we can do is measure how close a play is to a particular norm. Notice that players face a sequence of binary choices, that in turn can be viewed as a probability vector in a simplex. For the Principle, she has a single choice - how much bonus to pay when her observed signal is H or L. Hence, the strategy of the Principle is given by:

$$\sigma_P = (p_1, p_2) \in \Delta^2 = \{(p_1, p_2) \,|1 \geq p_1, p_2 \geq 0\},$$

where p_1 is the probability of paying a bonus when the signal is high ($s_P = H$), and p_2 is the probability of paying a bonus when the signal is low ($s_P = L$). With this notation it follows:

$$\sigma_P^{PI} = (1,1), \sigma_P^{PO} = (1,0), \sigma_P^{R} = (0,0).$$

The Worker has a more complex decision. He needs to chose high or low effort and then choose to initiate conflict depending upon his bonus and his observed signal. Suppose that p_E is the probability of high effort. Let his information set be given by I=$\{BH, BL, 0H, 0L\}$, where BH means a bonus was paid and he observes a high signal, $0H$ is no bonus and low signal and so on. Thus the strategy of the Worker can be given by vector of probabilities of effort and imposing a cost upon the Principal:

$$\sigma_W = (p_E, p_{BH}, p_{BL}, p_{0H}, p_{0L}) \in \Delta^5.$$

The three types of norm behaviors are:

$$\sigma_W^{PI} = (1,0,0,1,1),$$
$$\sigma_W^{PO} = (1,0,0,1,0),$$
$$\sigma_W^{SI} = (0,0,0,0,0).$$

Notice that the extensive form game tree is a binary tree with three players: Nature, Principal and Worker. Since it is a binary tree we can estimate the strategies using a multi-nomial logit. Specifically, each norm implies a binomial distribution at each node, which in turn implies a probability distribution over the final nodes in the game that can be represented by a logit model. We can then estimate how well the data is represented by a particular combination of norms. The details are provided in Section 4.

3. The Experiments

Our interest in this paper is to investigate the endogenous formation of an efficiency-enhancing conflict culture. In our first experiment (the baseline experiment) participants play the subjective evaluation game without any further intervention or clarification from the experimenter. We identify the causal effect of the opportunity to engage in conflicts by comparing a treatment in which conflicts are excluded by design (no conflict treatment) to a treatment in which the conflict option is available (conflict treatment). As we will illustrate in detail below, in this first experiment the participants failed to use the conflict option in a beneficial way so that efficiency was lower in the conflict treatment than in the no conflict treatment. In response to the results of the first experiment, we decided to run a series of additional experiments to explore different factors that potentially facilitate the emergence of a coordinated conflict norm. In total, we ended up running three additional experiments that explore the impact of structured communication (the communication experiment), codes of conduct (the agreement experiment), and an institutionalized procedures (the grievance experiment).

Each experiment stands on its own and includes at least 10 independent sessions comparing a conflict and a no conflict treatment. Each experiment is based on data from about 200 participants (in all four experiments together we had 852 participants in 42 sessions). As each experiment was conducted in a different time period, we first present the results of each experiment in isolation and do not pool our data in this section. We establish our main results using two-sided, non-parametric tests that use session-level averages from the same experiment as independent observations. In the final part of this section we also present additional results using a pooled data set that includes observations from all experiments together. Doing so allows us to apply more complex statistical models including regression analysis with standard errors clustered at the session level.

We next describe the subjective evaluation game underlying all our experiments and provide details regarding choice structure and parameters. Thereafter we describe each of our four experiments separately. In particular, for each experiment we outline the specific design characteristics, describe the procedures, and finally report the results. We wrap up the section with additional results based on the pooled data set including all experiments.

3.1. Implementation of the Subjective Evaluation Game.

All our experiments are based on the same implementation of the subjective evaluation game. In this section we outline a step-by-step account of the game and describe each player's strategy space and decision sequence in detail. Participants play the experiment for a total of 15 periods. Principals and workers are randomly re-matched at the beginning of every period. Each period consists of the following steps:

(1) *Matching and Contract*: Each worker is randomly assigned to a principal. The Worker gets a standardized contract. The contract guarantees a fixed wage $w = 100$, which the Worker gets with certainty. In addition, the Principal has the possibility to pay a bonus ($b \in \{0, B = 50\}$) after he has received a private signal s_P about the output r produced by the Worker.

(2) *worker's Effort Choice*: The Worker can choose between a low ($e = 0$) and a high effort level ($e = 1$). Choosing the high effort level is associated with a cost $c = 10$. The effort exerted by the Worker produces a return that can either be high ($r_H = 350$) or low ($r_L = 150$). The effort choice defines the probability with which the output is high or low: $Prob(r = r_H|e = 1) = Prob(r = r_L|e = 0) = 0.85$ and $Prob(r = r_L|e = 1) = Prob(r = r_H|e = 0) = 0.15$.

(3) *Output Determination and Signals*: After the Worker has chosen his effort level, a computerized random device determines the realized output. Neither the Principal nor the Worker observe the output. Instead, each one of them receives a private signal s indicating either a high output ($s = 1$) or a low output ($s = 0$). The Principal's signal depends on the realized output r: $Prob(s_P = 1|r = r_H) = Prob(s_P = 0|r = r_L) = 0.75$. The Worker's signal depends on the Principal's signal s_P: $Prob(s_P = 1|s_P = 1) = Prob(s_P = 0|s_P = 0) = 0.75$.[7]

(4) *Principal's Bonus Payment*: After the Principal observes the signal, s_P, she makes a decision regarding the bonus payment b.

(5) *worker's Initiation of Conflict*: Finally, the Worker observes the private signal s_A and the Principal's bonus choice b and decides to initiate conflict. When initiating a conflict the Worker determines the intensity of the conflict $d \in [0, 100]$. Conflict harms the Principal by reducing her payoff by $f(d) = dF = d$ and creates a cost $k(d) = dK$ for the Worker, where $K \in \{0.1, \infty\}$ depending on the treatment.

Payoffs. For the moment suppose that when the Worker decides to punish she sets $d = 1$, the maximal punishment. We can then compute the social surplus for each combination of the three strategies, self-interest, pay for input and pay for output. These are shown in Table (1).

[Table 1 about here.]

Notice that the pay for input has the highest surplus, followed by the pay for output. This is because it entail high effort, and no conflict costs. If workers and firms were trustworthy, then this would be the outcome all would agree upon. However, it is not stable. We can see this in the next figure which normalizes output to the equilibrium utility, and the gain from deviation. Thus movement along a column corresponds to changes in the Principal's payoff, while movement along a row corresponds to changes in the worker's payoff relative to the norm for that row.

[Table 2 about here.]

Observe that when strategies are restricted to one of the three norms of behavior, then both "self-interest" and *pay-for-output* are stable norms in the conflict treatments. Neither the Principal, nor the Agent gain from deviating from the norm. This is not true of the "pay for input" norm. The Worker has a net gain of 10 by moving to the self-interest norm. However, as we shall see, there is evidence that some players would like to coordinate upon the "pay for input" norm. The purpose of the experimental manipulations is to see if it is possible to have players coordinate upon a single "organizational culture". Under the no conflict treatment, "pay for output is no longer an equilibrium since deviations by the Principal are no longer punished. Total surplus for the Pay for Output norm in that case is the same as in the Pay for Input norm.

[7]Obviously, it would be possible to write down an equivalent structure of signals in which both signals depend on output. However, we chose this way of presenting the signals to the participants, because it makes the imperfect correlation of the two signals very transparent to the participants.

3.2. **Experiment 1: Baseline. Design**

In our first experiment the participants played the game exactly as described in section 3.1. We implemented the following two treatments:

- **No Conflict Treatment**: In this treatment the interaction of the Principal and the Worker ends after the bonus payment of the Principal. Formally, this corresponds to a version of the subjective evaluation game in which conflicts are prohibitively costly for the Worker, $K = \infty$. In this treatment Step 5 (the conflict stage) of the above procedure does not exist.
- **Conflict Treatment**: In this treatments all 5 steps of the game are played and the cost of conflict is $K = 0.1$.

Data Collection and Procedural Details

We conducted the study at the behavioral laboratory of HEC Lausanne (LABEX). Participants were recruited from the regular subject pool, covering all fields of study. We used ORSEE (Greiner, 2015) for the recruitment and z-Tree (Fischbacher, 2007) for programming the experiment. We ran 10 sessions with a total of 204 subjects and we conducted the sessions in November and December 2012. Sessions were randomly allocated to treatments and within sessions participants were randomly assigned to roles. We aimed at 24 subjects per session but some sessions were smaller due to no-shows. Five sessions had 16 subjects only and the other sessions had either 18 or 20 subjects. Sessions lasted for 50 to 80 minutes including the reading of the instructions and the final cash payments. Subjects received a show-up fee of 10 CHF and experimental points were converted at a rate of 70 points per CHF. Average total earnings were 23.8 CHF (24.2 CHF for subjects in the role of worker, and 23.3 CHF for subjects in the role of employer). The roles in the experiment were labeled as "worker" and "employer". Role assignments, choices, and earnings were completely anonymous.

Results

We first illustrate how the conflict treatment affects the main variables of interest (effort and surplus). Subsequently, we explore the underlying mechanisms by investigating how conflict strategies affect bonus payments and effort choices.

Result 1 ((Baseline - Effort and Surplus))**.** *The opportunity to engage in costly conflict leads to a small and statistically insignificant increase in worker's effort. However, conflict costs dominate efficiency gains from higher effort, so that total surplus in the conflict treatment is lower than in the no conflict treatment.*

The top row of Figure 5 compares the relative frequency of high effort (left-hand side) and the average total surplus (right-hand side) in the no conflict and conflict treatment of the baseline experiment. The Figure shows that the rate with which workers choose high effort rises from 40.4% in the no conflict treatment to 47.5% in the conflict treatment. This corresponds to an increase of 18%, but the increase in effort is not statistically significant (RS: $p = 0.420$).[8] Moreover, the figure also reveals that the increase in the frequency of high effort does not translate into a higher total surplus. On the contrary, total surplus decreases insignificantly from 233.8 in the no conflict treatment to 221.6 in the conflict treatment (RS: $p = 0.222$).

Figure 5 illustrates the reason for the lower surplus in the conflict treatment. The figure contrasts the total cost of conflict (i.e., the worker's cost of initiating the conflict plus the damage imposed on the Principal)

[8]If not explicitly stated otherwise reported p-values are based on two-sided, non-parametric rank sum (RS) or signed rank (SR) tests using session averages as independent observations.

with the additional gains from trade that result from the increased effort. Average conflict costs (20.0) are larger than average gains from higher effort (7.8) resulting in a negative effect of conflict on efficiency.

As our participants play the game for 15 periods with different partners, learning effects might occur and it is important to look at the dynamics. Table 3 presents the relative frequency of high effort and average total surplus over time (in bins of 5 periods). The table reveals that effort and surplus are subject to a downward trend in both treatments, but in particular in the absence of conflict. The frequency of high effort drops by 19% in the conflict condition (SR: $p = 0.188$) and by 35.6% in the no conflict condition (SR: $p = 0.063$). Total surplus decreases by 10% in the conflict condition (SR: $p = 0.063$) and by 13.4% in the no conflict condition (SR: $p = 0.063$). Although the presence of conflict slows down the decrease in effort provision over time compared to the no conflict treatment, the difference between effort levels across treatments remains insignificant, even in the final periods of the experiment (RS: $p = 0.310$). Moreover, total surplus keeps being superior in the absence of conflicts. Hence, result 1 holds throughout the experiment.

[Table 3 about here.]

To understand the reasons for why conflicts fail to be efficiency-enhancing, it is instructive to take a careful look at the conflict pattern that emerged in our first experiment.

Result 2 ((Baseline - Conflict Pattern))**.** *Most workers in the conflict treatment either refrain from engaging in conflict or follow a conflict pattern that aims at enforcing the pay-for-input norm. As a consequence, the worker's subjective performance signal has a very low impact on the conflict rate. If conflicts are initiated at all, they most frequently occur when workers provide high effort, but do not receive a bonus.*

The top row of Figure 5 presents the relative frequency of conflict initiation contingent on three determinants, the worker's effort level, the worker's subjective performance signal and the Principal's bonus payment. The left-hand panel covers low-effort observations, the right-hand panel covers high-effort observations. Two insights emerge from the observed pattern in this figure: First, the punishment decision of a large set of workers is guided by the pay-for-input norm: the conflict rate is highest if workers who decide to exert high effort are not rewarded with a bonus. Second, the pay-for-output norm matters only for a minority of workers as the subjective performance signal of workers has a weak impact on the decision to engage in conflict. We explain these two points in detail in the subsequent paragraph.

Figure 5 shows that the worker's effort choice in combination with a lack of a bonus payment is the most important determinant of conflict. The highest conflict rate is observed if the Worker chooses high effort, gets a positive subjective signal, but does not receive a bonus (64.4%). Keeping everything else constant (high effort, positive worker signal) the conflict rate drops to 10.7% if the Principal pays the bonus, a decrease of 83% (SR: $p = 0.063$). Likewise, if we compare the highest conflict rate (64.4%) to the corresponding low-effort case (keeping the absence of a bonus and the positive worker signal constant) the conflict rate drops to 26.1%, a decrease of 59.4% (SR: $p = 0.063$). At the same time, the figure also reveals that the worker's signal has a weaker impact on conflict initiation. If we compare the highest conflict rate (64.4%) to the case in which the worker's subjective signal is negative (keeping the high effort and the absence of the bonus payment constant), the conflict rate only drops from 64.4% to 48.6%, a decrease of 24.5% (SR: $p = 0.063$).[9]

[9]The finding that the conflict rate is predominantly determined by bonus payments and effort choices receives further support when we focus the analysis on the cases in which the Worker receives a negative signal. If the Worker exercises high effort, the conflict rate decreases from 48.6% to 10.7% if the Principal decides to pay the bonus (SR: $p = 0.063$). Similarly, if we keep the absence of a bonus payment constant, the conflict rate drops from 48.6% to 20.9% if the Worker decides to exert low effort (SR: $p = 0.125$).

Figure 5.4 provides a different perspective on workers' conflict strategies. In this figure, instead of averaging behavior across subjects, the figure shows different distributions of punishment patterns taking individual workers as the unit of observation, i.e., each dot in the figure represents a single worker. Moreover, as conflicts emerge predominantly when the Worker does not receive a bonus, the figure focuses only on conflict initiation in the cases in which the Principal decides not to pay a bonus. The figure illustrates the extent to which the decision to initiate conflict depends on the worker's subjective signal (given that no bonus payment has been made). On the x-axis (y-axis), we measure the probability of initiating conflict if the signal is negative (positive). In addition, the figure also reveals each worker's effort level: small (large) dots represent workers that mostly picked a low (high) level of effort. Finally, the figure uses colors to categorize the workers. We distinguish four types of workers: pay-for-output norm enforcers (PfI), pay-for-input norm enforcers (PfI), self-interest non-enforcers (R) and motivated non-enforcers (MNE). Pay-for-output norm enforcers pick high effort and only engage in conflict if they do not receive a bonus after having observed a positive subjective signal. These workers appear as large dots in the top-left quadrant. Pay-for-input norm enforcers also choose high effort, but they ignore their signal and engage in conflict whenever they do not receive the bonus. These workers appear as large dots in the top-right quadrant. Motivated non-enforcers are workers who exert high effort, but do not engage in norm enforcement (i.e., stay away from conflicts). They appear as large dots in the bottom-left quadrant. Rational non-enforcers, finally, pick low effort and do not engage in conflict. They appear as small dots in the bottom-left quadrant.

The figure shows that the worker's population fails to coordinate on any particular norm. To begin with, only 42.5% of all workers can be defined as norm enforcers (PfI or PfO). Moreover, there are more pay-for-input norm enforcers (25.5%) than pay-for-output norm enforcers (17.0%). The majority of non-enforcers (46.8%) are rational and exert mostly low effort (36.2%) and motivated non-enforcers are rare (10.6%). The remaining workers (10.6%) show profiles that do not correspond to either of the four types described above.

In Table 3 we present the development of the conflict rate over time. The table reports conflict initiations for cases in which no bonus has been paid contingent on the worker's effort level. Most interesting is the development of the conflict rate in the situations in which the Worker exerts high effort. The table reveals that in the first 10 periods of the experiment, the conflict rate is only slightly higher when the worker's subjective performance signal is positive than when it's negative and the difference is not statistically significant. In the final five periods, in contrast, we observe a substantial increase in the conflict rate when the worker's signal is positive (from roughly 60% in periods 1-10 to more than 80% in periods 1-15). No such increase is observed for the cases in which the worker's signal is negative. As a consequence, the conflict rate in the final five periods is marginally significantly higher for a positive worker signal than for a negative one (SR: one-sided $p = 0.063$). However, despite the fact that the dynamic analysis suggests that the pay-for-output norm may increase in importance over time, it is also important to emphasize that the conflict rate with negative signals remains roughly constant and at a high level throughout the experiment ($\tilde{4}$5-52%). This observation confirms that workers fail to coordinate on the enforcement of the pay-for-output norm (even after having played the game for many periods).

[Table 4 about here.]

From a theoretical point of view, the workers' failure to coordinate on the pay-for-output norm is problematic, because the somewhat erratic conflict patterns displayed in Figures 5 and 5.4 does not create well-aligned financial incentives for principals to use their bonus payments in a way that rewards those

workers who are perceived to be productive. A more detailed investigation of principals' bonus payments confirms the empirical relevance of this concern.

Result 3 ((Baseline - Bonus Payments)). *Principals in the conflict treatment pay the bonus significantly more often than principals in the no conflict treatment. However, the bonus payments in the conflict treatment are only partially consistent with the pay-for-output norm. Although the principals pay the bonus more often when their subjective signal is positive, bonus payments also occur quite frequently when the Principal's subjective signal is negative.*

The opportunity to engage in conflict has a strong positive impact on the overall frequency with which principals pay a bonus to the Worker: the bonus rate increases from 14.3% in the no conflict treatment to 42.6% in the conflict treatment (RS: $p = 0.008$). Figure 5.5 shows the frequency of bonus payments as a function of the Principal's private signal in both treatments. The pattern observed in the bonus payments is only partially in line with the pay-for-output norm; although in both treatments principals are more likely to pay the bonus if their subjective signal is positive, bonus payments occur frequently when the signal is negative. In the conflict treatment the bonus rate is 57.4% if the Principal's subjective signal is positive and 29.0% if the signal is negative (SR: $p = 0.063$). The corresponding numbers in the no conflict treatment are 24.7% and 5.4% (SR: $p = 0.063$).

Table 5 shows bonus payments in both treatments over time contingent on the signal observed by the Principal. In the conflict treatment the observed patterns remain stable over time. In the no conflict treatment bonus payments show a decreasing trend when the signal is high (this development occurs in parallel with the observed decrease in effort displayed in Table 3).

[Table 5 about here.]

Figure 5.6 displays distributions of bonus payment patterns in the conflict treatment at the individual level; each dot in the figure represents one principal. The only information that principals have when deciding about the bonus payment is their subjective performance signal. The Figure, therefore, displays the frequency with which principals pay the bonus as a function of their signal. The horizontal axis displays the bonus rate in response to a negative signal and the vertical axis the bonus rate in response to a positive signal. Similar to Figure 5.4, this figure allows to identify norm compliance, but this time for the principals instead of workers. Principals who follow the pay-for-output norm and pay bonuses predominantly after having observed a positive signal are located near the top-left corner. Those who subscribe to the pay-for-input norm and pay bonuses irrespective of their signal are located near the top-right corner. Principals who follow neither norm and never pay bonuses appear near the bottom-left corner.

The Figure reveals that no principal strictly sticks to a particular fairness norm. The large majority of principals pay the bonus more frequently after having observed a positive signal (55.1%), but there is considerable heterogeneity in both the bonus frequency and the weight that principals assign to their subjective signal. This suggests that worker's inability to coordinate on the enforcement of the pay-for-output norm implies is consistent with the observation that many principals also fail to coordinate on the norm. The lack of coordination on the side of principals, in turn, implies that the bonus payments fail to create a monetary incentive for workers to exert high effort. When comparing the profits of workers before the conflict stage, we find that the expected gains from exerting low effort in the conflict treatment amount to 19.1, whereas the expected gains from exerting high effort are only 14.1. The fact that the opportunity to engage in conflicts does not prevent self-interested workers from being better off by choosing low effort

provides an explanation for the high population fraction of rational non-enforcers observed in Figure 5.4 and the limited positive impact that the conflict treatment has on overall effort (see Result 1).

Our first experiment illustrates that it is not obvious that conflicts have an efficiency-enhancing effect in a work environment with subjective performance evaluation. Our results confirm that the presence of multiple fairness norms can lead to coordination failure which, in turn, implies that conflicts fail to have the desirable motivating effects. As a consequence, the cost of conflicts dominate and total surplus is lower in the presence of conflict opportunities than in their absence.

3.3. **Experiment 2: Communication. Design**

The lack of communication is arguably one factor that renders coordination difficult in our baseline experiment. There is ample evidence in the literature that pre-play communication helps to overcome coordination failure (see, e.g., Cooper et al., 1992; Brandts and MacLeod, 1995; Cason and Mui, 2007; Brandts and Cooper, 2007) and communication opportunities are available in most real-life settings in organizations. Our second experiment (subsequently termed "communication experiment") therefore investigates the extent to which communication helps the trading parties to coordinate on an efficiency-enhancing use of conflicts. The design of the experiment is identical to the one of the baseline experiment, but we now add a communication stage at the beginning of every period. In the communication stage both the Principal and the Worker could pick a message to their trading partner from pre-defined sets of messages.[10]

The message set of the Principal was the same in both the no conflict treatment and the conflict treatment. It contained the following announcements of different bonus strategies:

(1) I will pay the bonus with certainty.
(2) I will pay the bonus with high probability.
(3) I will pay the bonus if I have the impression that you exerted high effort.
(4) I pay the bonus if my private information indicates a high return.
(5) I will not pay the bonus.
(6) I prefer not to send a message.

In the no conflict treatment workers' messages contained only announcements of their effort strategies. The following options were available:

(1) I will exert high effort.
(2) I will exert low effort.
(3) I prefer not to send a message.

In the conflict treatment workers could add a second part to their message in which they announced their conflict strategies:

(1) I will always reduce the return.
(2) I will reduce the return if I do not get the bonus.
(3) I will reduce the return if I do not get the bonus although I got a good signal.
(4) I will never reduce the return.
(5) I prefer not to send a message.

[10]There is an emerging consensus in the experimental literature that in many environments free-form communication is more effective than structured communication (Brandts and Cooper, 2007; Brandts et al., 2015). Nevertheless, we consciously decided to use structured communication in this experiment, because including the different fairness norms in the set of pre-specified messages allowed us to make these norms very salient to the participants. This point is particularly relevant in light of the fact that the pay-for-output norm is somewhat complicated. Including it in the set of messages made this choice available even to participants who did not think about this possible norm themselves.

This set of pre-specified messages includes the fully rational strategy, the pay-for-input norm, the pay-for-output norm and the possibility not to communicate. Messages were selected and sent simultaneously.

Data Collection and Procedural Details

The laboratory, the subject pool, the recruitment process and the software used to program and run the experiment were the same as in baseline experiment. We ran 10 sessions with a total of 176 subjects. We conducted them in March, April, May and December 2013. We aimed at 24 subjects per session but some sessions were smaller due to no-shows. Five sessions had 16 subjects only and all other sessions had at least 18 subjects. Average total earnings were 35.8 CHF (36.6 CHF for subjects in the role of worker, and 35 CHF for subjects in the role of employer).

Results

We begin by analyzing how conflicts affect effort and surplus in the presence of communication before we turn to a detailed investigation of communication strategies, conflict patterns and bonus payments.

Result 4 ((Communication - Effort and Surplus))**.** *The impact of conflicts on efficiency in the presence of pre-play communication is similar to the one observed in the baseline experiment. The availability of conflicts leads to a slight increase in workers' effort, but the effect remains insignificant. Total surplus is slightly lower in the conflict treatment than in the no conflict treatment.*

The second row of Figure 5 displays the relative frequency of high effort and average total surplus in the communication experiment. The rate with which workers choose high effort rises from 51.8% in the no conflict treatment to 58.5% in the conflict treatment. This corresponds to an increase of 13%, but the effect is not statistically significant (RS: $p = 0.690$). Total surplus decreases insignificantly from 243.7 in the no conflict treatment to 237.7 in the conflict treatment (RS: $p = 1.000$). The top, right-hand panel of Figure 5 contrasts the total cost of conflict with the additional gains from trade that result from the increased effort in the conflict treatment. Average costs (22) are larger than average benefits (16) resulting in the above reported negative effect of conflict on surplus.

Table 6 presents effort and total surplus over time (in bins of 5 periods). The table reveals that the difference in effort arises exclusively in the final periods of the experiment. In periods 11-15 workers choose high effort in 57.8% of the cases in the conflict treatment, but only in 34.0% of the cases in the no conflict treatment. However, considerable variance across sessions implies that this difference remains statistically insignificant (SR: $p = 0.420$). Total surplus in the conflict treatment never surpasses total surplus in the no conflict treatment in any phase of the experiment.

[Table 6 about here.]

Result 4 reveals that the availability of structured pre-play communication does not trigger the efficiency-enhancing effects of conflicts. In the following we shed more light on the reasons for this result. We begin with an analysis of communication choices.

Result 5 ((Communication - Message Choices))**.** *Participants most frequently choose messages that are in line with the pay-for-input norm. First, a large majority of workers announce that they will exert high effort and threaten to engage in conflict if they do not get a bonus. Second, the majority of principals state that they will pay the bonus if they have the impression that the Worker exerted high effort, but only a minority communicates explicitly that they plan to make their bonus payment contingent on their subjective signal.*

We start with the communication strategies of workers. The vast majority of workers in both the conflict condition and the no conflict condition chooses to send the message: "I will exert high effort" (79.0% and

85.6%, respectively) and only a small proportion in both treatments chooses the message "I will exert low effort" (15.1% and 12.1%, respectively). Regarding conflict strategies the most frequent message chosen is "I will reduce the return if I do not get the bonus" (55.1%). This observation hints at the fact that many workers seem to subscribe to the pay-for-input norm. Instead, the message corresponding to the pay-for-output norm "I will reduce the return if I do not get the bonus although I got a good signal" was picked less than half as often (21.8%). Other messages played only a minor role ("I will never reduce the return", 16.7%) or barely any role at all ("I will always reduce the return", 0.3%).

Principals remain rather non-committal in their bonus payments announcements. In both the conflict condition and the no conflict condition, nearly half of the principals choose to send "I will pay the bonus if I have the impression that you exerted high effort" (49.9% and 49.6%, respectively). To interpret the choice of this message, it is important to keep in mind that principals also had the possibility to choose the message "I pay the bonus if my private information indicates a high return" (which was picked by 25.6% and 20.0%, respectively). Thus, principals who choose the former message decide explicitly not to signal that they will make their bonus payment strictly contingent on their subjective information because, otherwise they should have selected the latter message. From this point of view, it seems inconsistent to interpret the most frequently chosen message as being fully in line with the pay-for-output norm. In fact, the message can also be consistent with the pay-for-input norm, if the Principal plans to base her impression on the worker's effort message. The other messages available to principals were chosen with low frequency: "I will pay the bonus with certainty" (12.6% and 15.3%, respectively), "I will pay the bonus with high probability" (8.4% and 9.0%, respectively), and "I will not pay the bonus" (0.6% and 1.1%, respectively).

The analysis of the communication strategy suggests that most trading parties signal to each other that they intend to follow the pay-for-input norm. The most frequent message combination (26.2%) is the one in which the Worker announces to exert high effort and to engage in conflict if no bonus is paid and the Principal announces to pay the bonus as long as she has the impression that the Worker exerts high effort. Next we analyze the extent to which actual behavior corresponds with these messages.

Result 6 ((Communication - Conflict Pattern))**.** *Most workers in the conflict treatment either follow conflict patterns that are in line with the pay-for-input norm or abstain from engaging in conflicts. Conflict patterns consistent with the pay-for-output norm are rarely observed.*

The second row of Figure 5 presents the relative frequency of conflict initiation in the communication experiment contingent on the worker's effort level, the worker's subjective performance signal and the Principal's bonus payment. The panel on the right-hand side covers observations in which workers exert high effort. The conflict pattern observed in these situations is in line with the pay-for-input norm behavior: the conflict rate is high if the workers receive no bonus and low otherwise, moreover, the subjective signal of the workers plays no role. In the absence of a bonus a high signal triggers conflicts in 65.8% of the cases and a low signal in 66.1% of the cases (SR: $p = 0.625$). Somewhat surprisingly, we observe a similar pattern for observations with low effort levels: in the absence of a bonus a high signal triggers conflicts in 56.7% of the cases and a low signal in 51.9% of the cases (SR: $p = 0.625$).

To better understand the conflict strategies of workers it is instructive to take a look at Figure 5.4 which shows distributions of punishment patterns based on observations at the individual level (see Result 2 for a detailed explanation of this figure). The figure reveals a strong clustering of observations near the top-right corner. The cluster in the top-right shows both large orange dots, but also small red dots. Whereas the large orange dots correspond to individuals who subscribe to the pay-for-input norm, the red dots represent

workers who systematically initiate conflicts without providing high effort themselves. This indicates that the conflicts observed in Figure 5 after not obtaining a bonus payment, are not only a consequence of imperfect pay-for-input norm enforcers (who shirk on effort from time to time), but are also caused by workers who do not provide high effort and systematically initiate conflicts.

In this treatment, workers tend to coordinate on the pay-for-input norm more frequently than in the baseline. Overall, the population size of norm enforcers remains the same as in the baseline, however, the concentration of norm enforcers who follow the pay-for-input norm is even more pronounced. The overall population share of norm-enforcer (pay-for-input and pay-for-output norms together) amounts to 44.4%, which is almost identical to the corresponding rate in the baseline experiment: 42.5%. However, the fraction of pay-for-input norm followers corresponds to 80% of the norm enforcers in the communication experiment compared to 60% in the baseline experiment. Moreover, the population share of motivated non-enforcers remains roughly constant (11.1% (communication) vs. 10.6 (baseline)), the share of rational non-enforcers decreases from 36.2% (baseline) to 19.4% (communication). However, this decrease is misleading because it is almost fully compensated by an increase in the share of unclassified workers who mostly consist of the above discussed systematic conflict initiators who do not provide high effort.

In Table 7 we show how the conflict rate after not receiving a bonus develops over time. The table distinguishes between observations with positive and negative subjective signals of the Worker and high and low effort provision. The Table reveals the absence of a response of the workers to their private signal when deciding to initiate conflict. This finding reinforces the interpretation that workers predominantly focus on the enforcement of the pay-for-input norm in the communication experiment.

[Table 7 about here.]

The analysis of the communication strategies and the conflict patterns suggests that workers have a strong focus on the pay-for-input norm in the communication experiment. From the analysis of the communication choices we already know that most principals do not explicitly announce that they will make their bonus payment contingent on their subjective signal (as in compliance with the pay-for-output norm would require), but seem to prefer a non-committed message that promises a bonus payment as long as the Principal has the impression that the Worker worked hard. We now explore how all this translates into the principals' decision to pay bonuses.

Result 7 ((Communication - Bonus Payments))**.** *As in the baseline experiment, principals are more likely to make bonus payments in the conflict treatment than in the no conflict treatment. However, in line with their non-committal announcements, principals do not seem to follow either norm very systematically. Most principals condition their bonus payments to some extent on their signal of Worker performance, but many principals also pay bonuses with high frequency after a negative signal.*

At the average level, we observe that conflict increases the fraction of bonus payments and, moreover, the large majority of principals pay the bonus more frequently after having observed a positive signal. The second row of Figure 5.5 shows the frequency of bonus payments as a function of the Principal's private signal. In the conflict treatment the bonus payment rate is 87% if the Principal's subjective signal is positive and 50.3% if the signal is negative ($p = 0.063$). The corresponding numbers in the no conflict treatment are 36.8% and 6.2% ($p = 0.063$). These bonus patterns suggest that principals are not systematically following a norm.

At the individual level, we observe that most principals play a mix of the pay-for-input and the pay-for-output norm. The top-right panel of Figure 5.6 displays bonus payment patterns in the conflict treatment

at the individual level. If most principals were to follow the pay-for-input norm (pay-for-output norm), dots should be clustered in the top-right (top-left) corner. Instead we observe that most dots lie between the top-right and top-left quadrant, indicating that the principals are not systematically following either norm.

Table 8 shows bonus payments contingent on the signal observed by the Principal over time. In the conflict treatment, the bonus payment frequency remains very stable over time. In the no conflict treatment, in contrast, the frequency of bonus payments decreases quite drastically towards the end of the experiment.

[Table 8 about here.]

Our second experiment shows that communication does not solve the coordination problem of principals and workers. If anything, workers seem to have an even stronger focus on the pay-for-input norm than in the baseline experiment (this is reflected in both their communication strategies and their actual choices). In the conflict treatment, principals do not systematically follow a norm and, accordingly, their bonus payments fail to provide financial incentives for the Worker to work hard. The expected gains for the Worker after exerting high effort in the conflict treatment are only 27.5, whereas the expected gains after exerting low effort are 31.1. As a consequence, the results resemble those of the baseline experiment; conflicts fail to enhance cooperation and efficiency because the costs of conflict outweigh the benefits of additional effort.

3.4. **Experiment 3: Agreement (Code of Conduct). Design**

The first two experiments have relied only on self-coordination. The results indicate that people seem to have a tendency to focus on the simpler and arguably more intuitive pay-for-input norm. Unfortunately, this norm is not well-suited to increase cooperation in the subjective evaluation framework we are interested in. The fact that spontaneously emerging conflict cultures may be ineffective or even counterproductive suggests that there may be an important role for active management. Code of conducts are one way through which many organizations transmit values, guidelines and proper practices to their workforce. Our next experiment (henceforth called "the agreement experiment") therefore aims at studying the impact of an external appeal to coordinate on the pay-for-output norm. In this experiment we begin each period with a stage in which both the Principal and the Worker are asked to electronically sign a non-binding agreement in which they confirm their intention to follow the pay-for-output norm. Failure to sign the agreement by at least one party implies that trade does not occur and an unattractive outside option ($u_r = 40$) is implemented. The agreement specifies a simple code of conduct for the relationship. The suggested choices correspond to the strategies prescribed by the pay-for-output norm. In the no conflict treatment the Worker agrees to exert high effort, whereas the Principal agrees to pay the bonus if and only if her subjective signal is positive. In the conflict treatment, the code of conduct has an additional paragraph in which the Worker agrees to engage in conflict only if he does not receive a bonus although his subjective performance signal was positive. In the instructions all participants are made aware of the fact that the agreement is not binding, i.e., signing the code of conduct does not alter the strategy set available to the players. An example of the agreement in English can be found in the Appendix.

Data Collection and Procedural Details

The laboratory, the subject pool, the recruitment process and the software used to program and run the experiment were the same as in the previous experiments. We ran 10 sessions with a total of 200 subjects and we conducted them in November and December 2017. We aimed at 24 subjects per session but some sessions were smaller due to no-shows. Two sessions had 18 subjects only and the other session had either 20 or 22 subjects. Sessions lasted for 50 to 80 minutes including the reading of the instructions and the final cash payments. Subjects received a show-up fee of 10 CHF and experimental points were converted at a rate

of 70 points per CHF. Average total earnings were 34.7 CHF (33.3 CHF for subjects in the role of worker, and 36.1 CHF for subjects in the role of employer).

Results: This section presents the results of our third experiment. We first analyze how the need to sign a code of conduct before interacting with a trading partner affects the main outcomes of the experiment:

Result 8 ((Agreement - Effort and Surplus)). *The introduction of a code of conduct does not substantially change the impact of conflicts on effort and surplus. As in the baseline experiment, workers' effort is not significantly higher and total surplus is lower in the conflict treatment than in the no conflict treatment.*

The third row of Figure 5 shows the relative frequency of high effort and average total surplus in the agreement experiment. The figure shows that the rate with which workers choose high effort rises from 49.1% in the no conflict treatment to 52.4% in the conflict treatment. This corresponds to an increase of 6.7%, but this effect is far from significant (RS: $p = 0.842$). Moreover, total surplus decreases from 243.1 in the no conflict treatment to 224.5 in the conflict treatment (RS: $p = 0.151$). Figure 5 shows that the average cost of conflict (24.9) is higher than the average benefit from higher higher effort (6.3) resulting in a negative effect of conflict on efficiency.

Result 8 indicates that the code of conduct fails to have a positive impact on overall outcomes. To better understand why the code of conduct doesn't have an effect on effort provision, we analyze the participants' decisions in the different stages of the game in more detail.

Result 9 ((Agreement - Acceptance of agreement)). *In the no conflict treatment, nearly all participants accept the agreement; in the conflict treatment all participants accept the agreement.*

Overall, a total of 97.5% of the participants accept the agreement (95.1% in the no conflict treatment and 100% in the conflict treatment) and there is no time trend in the acceptance rate in either treatment.

Result 10 ((Agreement - Conflict Pattern)). *Signing a code of conduct does not coordinate workers on the pay-for-output norm. Norm enforcement is rather week in the agreement experiment in general, but among the workers who engage in norm enforcement, the clear majority follows the pay-for-input norm.*

Despite the fact that all workers who interact with a principal sign a code of conduct that instructs them to follow the pay-for-output norm, the conflict pattern looks almost identical to the one observed in the baseline experiment: conflict initiation is mostly determined by whether the Principal pays a bonus or not. The third row of Figure 5 presents the relative frequency of conflict initiation in the agreement experiment contingent on the worker's effort level, the worker's subjective performance signal and the Principal's bonus payment decision. The Worker's subjective signal has only a very small impact. If the Worker exerts high effort, observes a positive signal, but gets no bonus, conflicts are initiated in 54.7% of the cases. If everything remains constant, but the Principal decides to pay a bonus, the conflict rate decreases to 6.2% (a decrease of 88.7%, SR: $p = 0.063$). However, if we compare the initial situation to the same situation except that the workers now observes a negative signal the conflict rate goes down only to 43.3% (a decrease of 21%, SR: $p = 0.188$). This pattern is consistent with the pay-for-input norm.[11]

The punishment strategies of individual workers displayed in Figure 5.4 further reinforces that the code of conduct doesn't affect the decisions of the parties. The population is divided into non enforcers (50.0%) and norm enforcers (26.1%). The share of norm enforcers in this experiment is nearly half the size of what

[11]The pattern is similar for low effort choices. Low effort, no bonus and a positive signal yield a conflict rate of 27.4%. A bonus payment lowers the conflict rate to 5.1% ($p = 0.126$), but a positive signal only reduces it to 22.1% ($p = 0.844$).

we observe in the baseline experiment (42.5%) or in the communication experiment (44.4%). Among the non enforcers, roughly two thirds exert low effort (blue dots - 32.6%) and the rest exert high effort (black dots - 17.4%). The share of rational non enforcers is the same as the share of pay-for-input norm followers (orange dots - 17.4%), while only 8.7% of the workers subscribe to the pay-for-output norm (green dots). Moreover, 23.9% of the workers cannot be defined by any of our pre-defined norms (red dots) giving further support to the claim that the code of conduct fails to coordinate workers.

Despite its failure to coordinate workers (and the resulting erratic punishment pattern), we observe that the code of conduct helps to coordinate bonus payments of principals.

Result 11 ((Agreement - Bonus Payments))**.** *As in the baseline and the communication experiments, principals pay bonuses more often in the conflict treatment than in the no conflict treatment. Moreover, signing the code of conduct induces more principal to follow the pay-for-output norm, that is, to condition their bonus payments on their subjective performance signal.*

At the average level, principals pay bonuses more often in the conflict treatment than in the no conflict treatment. Overall, the frequency with which principals pay the bonus nearly doubles, from 23.5% in the no conflict treatment to 42.8% in the conflict treatment (RS: $p = 0.008$). Furthermore, in this experiment, the signal becomes relevant in the decision to pay a bonus. The third row of Figure 5.5 presents the bonus payments in the agreement experiment contingent on the Principal's private signal in both the conflict and no conflict treatments. The bonus patterns shown in this figure indicate that the principals pay attention to their own subjective signal when deciding about the bonus payment. In the conflict treatment, the bonus rate is 20.0% if the Principal's subjective signal is negative and 65.8% if the signal is positive (an increase of 69.6% SR: $p = 0.126$). The corresponding numbers in the no conflict treatment are 8.2% and 39.2% (an increase of 79.1% SR: $p = 0.062$).

At the individual level, signing the code of conduct induces more principals to follow the pay-for-output norm. The bottom left panel in figure 5.6 shows the individual patterns. The figure reveals that 77.6% (green dots) of the principals in the agreement experiment subscribe to the pay-for-output norm (compared to 55.8% in the baseline experiment and 44.4% in the communication experiment). However, a careful examination of the figure, shows that the green dots representing the pay-for-output principals are concentrated between the 0.50-0.75 range of the vertical axis (probability of bonus payment after a high signal). Hence, although the tendency is to pay the bonus after a positive signal, principals do not do this often enough to create financial incentives for workers to exert high effort: the worker's expected payoff from exerting low effort is still higher (17) than the one from exerting high effort (15.5).

Our third experiment shows that a code of conduct alone does not successfully coordinate the trading parties on the pay-for-output norm. Although nearly all participants accept the agreement, only the principals follow the pay-for-output norm more frequently; the workers, in contrast, mostly choose to either abstain from conflict initiation or to enforce the pay-for-input norm.

3.5. **Experiment 4: Grievance. Design**

Our third experiment explored the idea that organizations might channel the disciplining effect of conflicts by employing a code of conduct to explicitly define the situations for which conflicts are appropriate. In our final experiment we take this idea one step further and study the impact of an institutionalization of conflicts. Instead of simply providing behavioral guidelines, organizations can also design explicit procedures that allow their members to file formal complaints against each other if the organization's code of conduct has been

violated. The design of the experiment is identical to the agreement experiment, except for one important detail: instead of simply initiating conflicts directly, workers need to follow a formal grievance procedure if they want to sanction their principal. The grievance procedures requires the filing of a standardized report in which the Worker needs to explicitly confirm that no bonus has been paid although the Worker has received a positive subjective signal. Once the report is submitted the punishment of the Principal is initiated with the exact same consequences as in the previous experiments. From an experimental point of view the institutionalization of the conflict corresponds to a pure re-framing of the sanctioning decision. However, the fact that the Worker needs to file a formal complaint implies that the Worker needs to lie in order to trigger a sanction in a situation in which the code of conduct does not deem a conflict appropriate. If workers are averse to lying, such a grievance procedure might help to coordinate workers on the pay-for-output norm (see, e.g., Gneezy, 2005; Lundquist et al., 2009; Gneezy et al., 2013).

Data Collection and Procedural Details

The laboratory, the subject pool, the recruitment process and the software used to program and run the experiment were the same as in the previous experiments. We ran 12 sessions with a total of 272 subjects and we conducted them in May 2017. We aimed at 24 subjects per session but some sessions were smaller due to no-shows. Two sessions had 18 subjects only and the other session had either 20 or 22 subjects. Sessions lasted for 50 to 80 minutes including the reading of the instructions and the final cash payments. Subjects received a show-up fee of 10 CHF and experimental points were converted at a rate of 70 points per CHF. Average total earnings were 35.6 CHF (33.8 CHF for subjects in the role of worker, and 37.3 CHF for subjects in the role of employer).

Results

We first explore how the introduction of a grievance procedure in addition to the code of conduct shapes the impact of conflicts on effort and surplus.

Result 12 ((Grievance - Effort and Surplus))**.** *In the presence of a grievance procedure, the opportunity to engage in conflicts has a significantly positive effect on workers' effort and leads to higher total surplus.*

The fourth row of Figure 5 displays the relative frequency of high effort and average total surplus in the grievance experiment. The figure shows that the rate with which workers choose high effort rises from 44.1% in the no conflict treatment to 64.4% in the conflict treatment. This corresponds to a statistically significant increase of 46% (RS: $p = 0.026$). The increase in effort in the conflict treatment also induces an increase in total surplus from 237.3 in the no conflict treatment to 243.7 in the conflict treatment. However, the increase in surplus effect is not statistically significant (RS: one-sided $p = 0.844$).[12]

An analysis of the dynamics reveals that institutionalized conflicts help sustain a higher effort level in all phases of the experiment. Table 9 presents the relative frequency of high effort and average total surplus over time. In the conflict condition, effort decreases only slightly over time, from 71.1% in the first five periods to 61.3% in the last five periods (a decrease of 13.8%, SR: $p = 0.063$). In the no conflict condition, in contrast, effort clearly declines over time, from 57.1% in periods 1-5 to 35.6% in periods 11-15 in the no conflict treatment (a decrease of 35.9%, SR: $p = 0.031$). Average surplus shows similar time patterns.

[Table 9 about here.]

[12]See the bottom right panel of Figure 5 for an illustration of the total cost of conflict (20.2) in comparison with the benefits of additional effort (26.2).

We next explore how the grievance procedure affects behavior in different stages of the subjective evaluation game.

Result 13 ((Grievance - Acceptance of agreement))**.** *In the no conflict and conflict treatments nearly all participants accept the agreement.*

Similar to the agreement treatment, most of the participants (98.8%) accept the agreement (98.1% in the no conflict treatment and 99.5% in the conflict treatment). Acceptance rates do not change substantially in either treatment over time.

Result 14 ((Grievance - Conflict Pattern))**.** *workers in the conflict treatment of the grievance experiment are more likely to follow the pay-for-output norm than the pay-for-input norm. As a consequence, the worker's subjective performance signal has an impact on the conflict rate. However, also in the grievance experiment there is a substantial share of workers who refrain from initiating conflict and do not enforce any norm.*

The fourth row of Figure 5 presents the relative frequency of conflict initiation in the grievance experiment contingent on worker's effort level, the worker's subjective performance signal and the Principal's bonus payment. We observe that, as in all previous experiments, conflicts are most often initiated when effort is high, the worker's signal is positive and no bonus is paid. In this case the conflict rate amounts to 61.2%. If all remains equal except that the Principal pays the bonus, the conflict rate decreases substantially to 1.9% (a decrease of 98%, SR: $p = 0.032$). However, whereas the impact of the bonus payment is comparable to what we observed in previous experiments, we now also observe that the worker's subjective signal has a decisive impact. If we replace the positive signal in the initial situation with a negative one, the conflict rate drops to 35.6% (a decrease of 42%, SR: $p = 0.032$).[13]

At the individual level, we observe that while institutionalized conflicts in the grievance experiment help to coordinate workers on the pay-for-output norm, it is not the case that all workers respect the code of conduct. The individual conflict patterns illustrated in the bottom-right panel of Figure 5.4 shows that, overall, 42.7% of the workers engage in norm enforcement. Two thirds of the norm enforcers follow the pay-for-output norm (27.9%), whereas the rest (14.8%) enforces the pay-for-input norm. Moreover, about a third of the worker's population (34.5%) abstains from initiating conflicts. Table 10 presents the development of the conflict rate of workers who do not receive a bonus over time.

[Table 10 about here.]

In the next result we explore the principals' bonus payments.

Result 15 ((Grievance - Bonus Payments))**.** *The structure of principals' bonus payments in the grievance treatment is similar to the one in the agreement treatment. Principals pay bonuses more often in the conflict treatment than in the no conflict treatment. Moreover, many principals follow the pay-for-output norm and condition their bonus payments on their subjective performance signal.*

The fourth row of Figure 5.5 shows the frequency of bonus payments in the grievance experiment as a function of the private signal of the Principal in both treatments. In this experiment bonus payments in the conflict treatment are highly contingent on the signal observed. The bonus rate is 79.3% if the Principal's subjective signal is positive and 20.2% if the signal is negative (SR: $p = 0.032$). A similar pattern (although

[13]The pattern is similar for low effort choices. Low effort, no bonus and a positive signal yield a conflict rate of 40.5%. A bonus payment lowers the conflict rate to 1.1% ($p = 0.062$) and a positive signal reduces it to 15.3% ($p = 0.094$).

at a lower level) is observed in the no conflict treatment where the corresponding numbers are 36.1% and 4.5% (SR: $p = 0.032$).

At the individual level, the bottom-right panel of Figure 5.6 adds supportive evidence to the claim that most principals in the grievance treatment condition their bonus payments on their subjective signal: 84.4% of the principals are classified as pay-for-output norm followers (compared to 55.1% in the baseline) and only 10.2% of principals follow the pay-for-input norm. These bonus patterns finally create (weak) monetary incentives for the Worker to exert high effort as the expected payoffs of exerting high effort are 20.1 and the gains from exerting low effort are 19.7.

In terms of bonus payment dynamics, table 5 show bonus payments over time contingent on the signal observed by the Principal and conflict. The bonus payment patterns remain very stable over time.

[Table 11 about here.]

Our final experiment shows that conflicts can have beneficial effects on motivation (and efficiency) even in a challenging environment like our subjective evaluation game. At the same time, the experiment also demonstrates that the establishment of a conflict culture is a delicate and somewhat risky endeavor. Despite the establishment of a clear code of conduct and grievance procedure that forces workers to lie if they wish to deviate from the code of conduct, not all workers follow the code of conduct and conflicts still emerge in situations in which they have no beneficial effects. These findings suggests that conflict cultures need to be carefully managed in order to be successful.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

4. Measuring Norms and Cultures

In our experiments the relationship between action and payoffs is unchanged across treatments. Nevertheless, we find that behavior is sensitive to context and that contracts make a difference. These findings are incompatible with models that assume fixed preferences over distributions of payoffs, but they are consistent with the idea that people may care about being behaviorally consistent with context-driven and endogenously selected norms. Our results illustrate that–if a group (or an organization) manages to successfully set up as set of interlocking norms–conflicts may not only increase the power of the worker, but also increase *effort*. We call such a set of interlocking norms a "culture".[14]

From an empirical point of view, our approach raises the important question of how to best measure the extent to which behavior is consistent with norm-abidance in such markets. In this section we provide a framework, using off-the-shelf econometric packages, that allows us to measure norm and cultural compliance in a simple manner. We build upon the fact that any strategic situation can be represented as an extensive form. We know that the end nodes of the decision tree provide a complete representation of the game. More precisely, let Z be the end nodes of an interaction between a Principal, a Worker and Nature (in game theory Nature is represented as a player with fixed mixed strategies). The outcome of an interaction induces a probability distribution over Z. A set of norms or culture effectively says that certain behaviors imply that some actions should never be taken, which in turn implies that there is a $Z^{0K} \subset Z$, where K refers to culture $K \in \{R, PI, PO\}$, such that $Pr\left[z \in Z^{0K}\right] = 0$. For example, under culture R (*self-interest choice*), effort should be zero ($E = 1$). Thus, if we observe many pairs with positive effort ($E = 1$), we get $Pr\left[Z^{0R}\right] > 0$, which implies that behavior is not well represented by rational choice theory.

However, our aim in this paper is obviously not to reject the rational choice model. Such findings are neither new, nor particularly useful. Rather, the question we want to ask is: which culture best fits our data? It might be the case that even though mistakes are frequent, rational choice still provides the best description of behavior (a common claim). The goal of this section therefore is to provide a method to measure the descriptive quality of any behavioral norm. Second, we use the term "culture" to mean a set of *interlocking norms*, so that we can also ask whether the population as a whole can be characterized by a common culture.

Our approach is inspired by Selten (1975)'s notion of a perfect equilibrium. He envisions players who at each node may make mistakes (trembling hand). In our experiment the game can be represented as a binary tree, and hence at each decision node choice can be viewed as a binomial distribution which can be represented using the logit model. For each context we can measure the distance between observed choices and a norm. Norm compliance in a particular context is less general than what one means by a "culture", which is a collection of norms that work together.[15] In our case, the *pay-for-output norm* is a collection of choices–high effort, followed by bonus if and only if the Principal observes a high signal, while the Agent initiates conflict if and only if there is no bonus and a high signal. After discussing norm compliance by the Principal and Worker, we measure the extent to which the population (organization) as a whole can be characterized by a single culture.

[14]This definition is consistent with a literature in economics that views culture through the lens of game theory. Kreps (1990) and Greif (1994) view culture as a Nash equilibrium to an employment game. MacLeod and Malcomson (1989, 1998) show that in a market context there can exist multiple self-enforcing "social norm" equilibria.

[15]There is a massive literature on culture. On the notion that culture is based upon constituent parts see Homans (1961).

The Principal's Norms. Let us begin by measuring the Principal's adherence to a norm. Figure (5.7) illustrates the Principal's choices parameterized for the case of the *pay-for-output norm*. The Principal cannot observe whether or not the Worker has produced effort, and hence the Principal does not know the probability of a high signal. The choice of a high signal is an action chosen by Nature that we can estimate as a binomial distribution characterized by γ_H that determines the probability of a high signal via:

$$Prob\left[s_P = H\right] = \frac{exp\left(\gamma_H\right)}{exp\left(\gamma_H\right) + exp\left(-\gamma_H\right)}. \tag{4.1}$$

In the figure we use the terminology $P \sim exp\left(\gamma_H\right)$ to indicate that the probability varies with γ_H modulo a factor to normalize to a probability as we have done in (4.1) (see DeGroot (1972) for details on this approach). Given the state, the Principal then chooses to pay a bonus or not. Notice that as $\gamma_{PO} \to \infty$, then

$$Prob\left[b = B | s_P = H\right] = Prob\left[b = 0 | s_P = L\right] \to 1.$$

These choices are exactly the ones indicated by the *pay-for-output norm*. The outcomes $z \in Z_P = \{Z1, ..., Z4\}$ are observed in the data. Since the realization of the signal is independent of the Principal's choice, then the *pay-for-output norm* described in Figure (5.7) satisfies:

$$Prob\left[z\right] = logit\left(X_z^{PO}\vec{\gamma}_{PO}\right), \tag{4.2}$$

where X_z^{PO} is defined in the figure and:

$$\vec{\gamma}_{PO} = \begin{bmatrix} \gamma_H \\ \gamma_{PO} \end{bmatrix}.$$

[Figure 7 about here.]

This model can be estimated from the data as a multi-nomial logit from which we obtain parameters $\vec{\gamma}_{PO}^* = (\gamma_H^*, \gamma_{PO}^*)$. If the Principal were following the PO norm perfectly, then $\gamma_{PO}^* \to \infty$. However, as we discussed above, we never get perfect compliance with any norm in any treatment, and hence these models are always well identified. A larger γ_{PO}^* indicates a behavior that is more closely approximated by the *pay-for-output norm*. In a similar way we can estimate γ_{PI}^* and γ_R^* for the *pay-for-input norm* and *self-interest choice* respectively.

Different norms correspond to different values for X_z, whose coefficients correspond to the strength or importance of a norm. The values of X_z for different norms are illustrated in Table (12).

[Table 12 about here.]

Notice that *rational choice* is a strategy that calls for never paying a bonus, and hence it is simply the opposite of the *pay-for-input norm*, which implies $\gamma_R^* = -\gamma_{PI}^*$. Thus we do not need to run separate regressions for *self-interest choice*. The results from estimating model (4.2) for each treatment are given in Table (13).

[Table 13 about here.]

The results confirm what we see visually in Figure (5.7). In the no conflict treatments the coefficient for the *self-interest choice* norm–as measured by $\gamma_R^* = -\gamma_{PI}^*$–is larger in value than the PO coefficient γ_{PO}^*. The negative coefficients mean that the Principal is more likely not to pay the bonus regardless of the signal. The fact that $\gamma_{PO}^* > 0$ for the no conflict treatment is a reflection of the frequency at which the low signal is observed by the Principal. In the no conflict treatments effort is lower and hence the low signal is observed

more often, so that there is a greater than 50% chance of the Principal observing the low signal, and then paying no bonus.

Things are very different in the conflict treatments. In the communication treatment the *pay-for-input* coefficient is indistinguishable than the *pay-for-ouput* coefficient. Since the *pay-for-input* coefficient is significantly positive, this implies that this norm dominates the *self-interest choice norm*, in contrast to the no conflict treatment where *self-interest choice* is a better fit. Hence the threat of conflict greatly increases the probability that the Principal pays the bonus.

In the communication treatment *pay-for-input* fits better than *pay-for-output*, but this result changes in the code of conduct and grievance cases. There the *pay-for-output norm* better describes the behavior of principals than either the *pay-for-output norm* or *self-interest choice*. Finally notice that the values for the code of conduct and grievance treatments in the no conflict treatment are comparable. Since there is no difference in payoffs between these treatments, and grievance has no real consequence, then this shows that the effect of "framing" is slight. Moreover, these results show that even though the experiments were done at different time, the results can be expected to be consistent across sessions. When the grievance procedure has a real effect, then there is a large and significant change in γ^*_{PO} for the two cases.

These comparisons are based upon the coefficients. We can also compare the overall fit of the models. The logit reports the value of the likelihood function, which is a measure of the overall goodness of fit, and provides us with a metric to rank models. Since each model corresponds to a different set of cultural norms, this approach provides a way to measure how closely a population of individuals comply with cultural norms. Since the models are not nested we can use the non-nested likelihood ratio test developed by Vuong (1989) to see if the overfits of the models differ significantly.

[Table 14 about here.]

In the no conflict treatments the R norm fits better than the PO norm in all treatments. These observations are confirmed by the Vuong test, and so we do not report those results here. However, matters are reversed with conflict. These results are illustrated in Table (14). The log-likelihood values come from the logit estimates. These provide a measure of the quality of the fit–bigger is better. The Vuoug test provides a way to compare two non-nested models. The null hypothesis is that the models are indistinguishable. The null is rejected when one model fits better than the other.

The *pay-for-input norm* and the *self-interest choice norm* are the same model but with $\gamma^R_P = -\gamma^{PI}_P$. For the treatments baseline, Code of Conduct and Grievance the null is rejects in favor of the the *pay-for-output norm*. In the case of communication the null is not rejected, which implies neither norm provides a superior fit to the data. The next issue is the extent to which the workers adjust their behaviors as a function of the treatments.

The Worker's Norms. Consider now the case of the Worker. In this case there are more information sets–the Worker observes both his signal and whether she gets a bonus. Moreover, the probability of a bonus is likely to be correlated with the signal, though the signal can be assumed to be exogenous as long as we condition on worker effort. Hence, in this case we have $\vec{\gamma}^*_{norm} = [\gamma^*_H, \gamma^*_{BL}, \gamma^*_{BH}, \gamma^*_{norm}]^T$, where γ^*_H measures the probability of a high signal, γ^*_{BL} measures the probability of a bonus given a low signal, γ^*_{BH} measures the probability of a bonus given a high signal, and, finally, γ^*_{norm} measures the probability of the worker following the $norm \in \{PO, PI, R\}$. As before, we run a multinomial logit model with standard errors clustered at the session level:

$$Prob\,[z] = logit\,(X_z\vec{\gamma})\,,$$

where $z \in Z_{Worker}$ and X_z are defined in Table (15).

[Table 15 about here.]

Workers know the effort level they have chosen, and so the question we ask is what norm does a worker who has high or low effort adopt. The results of the estimation are shown in Table (16). In the low effort case the workers are shirking, and hence presumably do not expect a bonus for their efforts. In this case, except for the Communication treatment, *self-interest choice* does fit better. However, it is not perfect, implying that many individuals who choose low effort also pay the cost of punishing the Principal who does not pay a bonus. Since effort is zero, this does not fit with a theory of reciprocity, but does suggest that there are individuals who shirk and get utility from punishing others who do not pay them. This effect is particularly prominent in the communication treatment.

In the case of high effort with baseline experiment all three norms have similar coefficients. With communication we see that the *pay-for-input norm* has the largest weight. The *pay-for-output norm* provides the best fit for both the code of conduct treatment and the grievance treatment, with the grievance treatment having a particularly large effect, going from 0.540 to 0.811. This result highlights the fact that Principal and Worker have different concerns. Above we saw that the code of conduct treatment did push the principals towards the *pay-for-output norm.*, but here we see that the effect on workers is more muted. Adding a formal grievance procedure provided for many individuals a mechanism that appears to legitimize imposing a cost upon a principal.

For the communication treatment the *pay-for-input norm* provides the best fit to the data, suggesting that in the absence of other coordinating information individuals appear to focus upon reciprocal norms, even though they are not self-enforcing.

[Table 16 about here.]

Next we compare the extent to which the population of workers conform to a particular social norm using the Vuong test. In this case we have three models, rather than two as in the Principal case. The Vuong test allows for only two models. We deal with this by comparing models to the top ranked model (and hence the top ranked model gets compared to the second ranked, and not the third). In the zero effort case *self-interest choice* is the dominant norm, and so again we do not report those results here. The effects of the treatments are reported in Table (17).

Under baseline the best fit is PO, while the second best is PI. The next column reports the test of the null hypothesis that norms R and PO are indistinguishable relative to norm R. Given the p-value of 0.998 we cannot reject the null. Next norm PI is compared to the null hypothesis that PO and PI are indistinguishable and again we cannot reject the null. Finally, PO, the best fitting norm is compare to the null that R and PO are indistinguishable and again the null is not rejected. Thus, in the absence of any coordination there does not seem to be convergence upon any norm

[Table 17 about here.]

In the case of the communication treatment the PI norm provides the best fit (as measured by the log likelihood), followed by the PO norm. We reject the null hypothesis that PO and PI are indistinguishable relative to the PI norm at the 0.5% level! Both the PO norm and R norms are rejected as potential best fits. However, even though there is communication, the results for the Principal show that they do not adopt *pay-for-input*. In this case there is a clear incentive for the workers to cheat, with the consequence that Principals often deviate from paying a bonus when they receive the low signal.

The interesting result is that under the code of conduct, in contrast to the results for the principal, no norm dominates. Finally, we find that in the case of the grievance treatment we can reject the hypothesis that PO and the second best norm, R, are indistinguishable relative to PO alone. Though the fit is better the likelihood function is quite small relative to the other cases indicating that there is still quite a bit of noise in workers' choice.

Culture. In this section we apply the same methodology, except now to the full population. The number of end nodes for the full game, denoted by Z, is $32 = 2^5$. These are found by combining all the possible binary choices: effort ($E \in \{L, H\}$), Principal's signal ($s_P \in \{L, H\}$), Agent's signal ($s_A \in \{L, H\}$), bonus pay ($bonus \in \{0, B\}$) and conflict ($pun \in \{Dont - Punish, Punish\}$). We do not need to consider the effect of productivity given by r since it is not observed by parties. The effect it has upon the signals is estimated endogenously. Thus, we run a model of the form:

$$Prob\,[z] = logit\left(\gamma^{norm} X_z^{norm} + \sum_{E\in\{0,1\}} \gamma_{Es_P} X_z^{Es_P} + \sum_{E\in\{0,1\},b\in\{0,B\}} \gamma_{Ebs_A} X_z^{Ebs_A}\right). \tag{4.3}$$

The first term estimates the weight, γ^{norm}, for the cultural norm. If all members of the population follow the norm perfectly then $\gamma^{norm} \to \infty$. Since perfection is not attainable, this value is always finite. Its size can be used to see how closely the population follows a particular norm. The remaining terms estimate the behavior of Nature in this relationship. Since the random draws are independent, these terms are unbiased estimates of the probabilities of the signals Agents observe. Table (18) reports the estimate of γ^{norm} for each conflict treatment. In the absence of communication, the *self-interest choice* culture provides the best description. When there is communication the *pay-for-input* culture dominates, followed by the *pay-for-output culture*. For the next two treatments the *pay-for-output culture* provides the best description of the culture in this laboratory experiment.

[Table 18 about here.]

We can get a sense of the quality of the fit by formally testing one culture against the other using a non-nested hypothesis test, as we have done above. These results are reported in Table (19). In contrast to the case of norms, we have a clear winner for each treatment. It is worth highlighting the point that we are estimated the same model for each treatment. Notice that in the Grievance case the *pay-for-output* culture best describes the behavior relative to the other cultures, however the likelihood values are smaller than in the other cases, which indicates that there is a great deal of noise in individual behavior.

[Table 19 about here.]

5. Discussion

In this experiment we show that in addition to affecting the allocation of resources, potential conflict can also increase overall performance. Potential conflict provides a self-help mechanism to implement agreements. This general point has been made many times within the context of the repeated prisoner's dilemma where conflict is represented by playing "cheat" for several periods. In those games cheating is perfectly observable, and hence it is clear when breach of an agreement has occurred. However, in our experiment a Worker who receives a good self evaluation of performance is not sure whether or not the Principal has received a similar signal. These creates ambiguity regarding what is "fair" behavior, and what is a "breach" of an agreement for which the Worker should impose a cost upon the Principal.

To analyze the stability and effectiveness of different plausible fairness norms in our setting, we introduce the notion of a "norm equilibrium" at which parties consider bundles of behaviors. We show that in our principal-worker relationship with potential conflict there are two norm equilibria, rational play and pay-for-output. The pay-for-output equilibrium provides higher payoffs than the rational play equilibrium, and hence it is a reasonable hypothesis that either party would play this equilibrium, or if permitted to communicate would recommend this equilibrium.[16]

It turns out that both of these hypotheses are false. In the baseline experiment *self-interest choice* provides the best, but imperfect fit to the data, while in the communication experiment parties tended to coordinate upon pay-for-input. However, norm compliance is very imperfect, and there are still more than 40% of the pairs in the communication treatment in which the worker chooses low effort. We do know that context matters. For example, Cohn et al. (2014) find that Bank employees are honest or not depending upon the work culture, and hence honesty is not a person specific immutable trait. After having observed the surprising results of our first two experiment, we therefore reasoned that the parties needed more nudging towards the pay-for-output equilibrium. In our third experiment parties enter into an explicit code of conduct agreement. We find that this additional element does move the culture towards pay-for-output to some extent (mostly on the side of principals), but average effort does not rise. The final experiment formalizes employment expectations and makes violations of the pay-for-output norm a "breach" event that gives the right to the worker to use a grievance procedure to complain about his treatment. This results in a significant increase in effort, and suggests that contracts, and formal procedures in an organization have a role to play in improving performance.

One of the most important implications of the analysis is to highlight some of the weaknesses of the principal-agent approach to contract design.[17] The focus in principal-agent theory is upon the quality of information and how it impacts contract design. It correctly shows that low-quality information can led to lower performing contracts. However, in practice what is typically emphasized is the need to link compensation to performance, which in turn has lead to many examples of dysfunctional employment relationship.[18] Our results show that another important ingredient to high performance entails providing formal power and voice to employees. Firms are often reluctant to do that because of the fear that it will result in rent seeking and lower performance. Our results do not dispute that observation, but suggest that with appropriate design, providing worker with more power in an organization can lead to enhanced performance. However, it is a delicate balance. Our framework provides a way to explore this question experimentally, and provides a way to quantify "culture" in a simple manner.

The term "culture" has many different meanings (Alesina and Giuliano (2015)). This experimental setting allows one to have a crisp definition of "culture" as a collection of interacting norms of behavior, where "norm" refers to behavior by a single individuals, broadly consistent with the notion of culture as used by Kreps (1990) and Greif (1994). One of the implications of our analysis is even though the material payoffs remain fixed across our experiments, we never the less observe a great deal of variation in behavior and performance. This may help explain some of the heterogeneity in organizational performance, even for firms in the same industry (Syverson (2011) and Bloom et al. (2012)). Obviously, this paper is only the beginning of a research agenda. We need much more work exploring different parameter values and

[16]The idea that parties would play the most efficient Nash equilibrium goes back to the work by Schelling (1980).

[17]See Gibbons (1997); Eisenhardt (1989) for thought assessments of the strengths and weaknesses of agency theory for understanding organizations.

[18]See Kerr (1975) and Hall (2000).

different organizational reforms to eventually better understand the full role of culture in organizational performance.

References

Akerlof, G. A. (1980, June). A theory of social custom of which unenployment May be one consequence. *Quarterly Journal of Economics 94*, 749–775. 1, 2

Akerlof, G. A. and R. E. Kranton (2000, Aug). Economics and identity. *Quarterly Journal of Economics 115*(3), 715–753. 1

Alesina, A. and P. Giuliano (2015). Culture and institutions. *Journal of Economic Literature 53*(4), 898–944. 5

Ambrus, A. and B. Greiner (2012). Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review 102*(7), 3317–32. 2

Axelrod, R. (1981, June). The emergence of cooperation among egoists. *American Journal of Political Science 75*(2), 306–318. 2

Binmore, K., J. Swierzbinski, S. Hsu, and C. Proulx (1993). Focal points and bargaining. *International Journal of Game Theory 22*(4), 381–409. Binmore, k swierzbinski, j hsu, s proulx, c. 1

Bloom, N., C. Genakos, R. Sadun, and J. V. Reenen (2012, February). Management practices across firms and countries. NBER Working Papers 17850, National Bureau of Economic Research, Inc. 5

Brandts, J. and D. J. Cooper (2007). It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association 5*(6), 1223–1268. 3.3, 10

Brandts, J., M. Ellman, and G. Charness (2015). Let's talk: How communication affects contract design. *Journal of the European Economic Association 14*(4), 943–974. 10

Brandts, J. and W. B. MacLeod (1995). On the strategic stability of equilibria in experimental games. *Games and Economics Behavior 11*, 36–63. 3.3

Cason, T. N. and V.-L. Mui (2007). Communication and coordination in the laboratory collective resistance game. *Experimental Economics 10*(3), 251–267. 3.3

Chaudhuri, A. (2011, Mar). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics 14*(1), 47–83. 1

Cohn, A., E. Fehr, and M. A. Marechal (2014, DEC 4). Business culture and dishonesty in the banking industry. *NATURE 516*(7529), 86–U190. 5

Coleman, P. T., M. Deutsch, and E. C. Marcus (2014). *The handbook of conflict resolution: Theory and practice*. John Wiley Sons. 1

Cooper, D. and J. Kagel (2016, 01). Other regarding preferences: A selective survey of experimental results. *The handbook of experimental economics 2*. 1

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1992). Communication in coordination games. *The Quarterly Journal of Economics 107*(2), 739–771. 3.3

De Dreu, C. K., M. J. Gelfand, et al. (2008). *The psychology of conflict and conflict management in organizations*. Lawrence Erlbaum Associates New York. 1

DeGroot, M. H. (1972). *Optimal Statistical Decisions*. New York, NY: McGraw-Hill Book C. 4

Dreber, A., D. G. Rand, D. Fudenberg, and M. A. Nowak (2008). Winners don't punish. *Nature 452*(7185), 348. 1

Egas, M. and A. Riedl (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences 275*(1637), 871–878. 1

Eisenhardt, K. M. (1989, jan). Agency theory: An assessment and review. *The Academy of Management Review 14*(1), 57–74. 17

Falk, A., E. Fehr, and U. Fischbacher (2003). On the nature of fair behavior. *Economic Inquiry 41*(1), 20–26. 6

Fehr, E. and U. Fischbacher (2003). The nature of human altruism. *Nature 425*(6960), 785. 1

Fehr, E. and U. Fischbacher (2004). Social norms and human cooperation. *Trends in cognitive sciences 8*(4), 185–190. 1

Fehr, E., O. Hart, and C. Zehnder (2011). Contracts as reference points: Experimental evidence. *American Economic Review 101*, 493–525. 1

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics 114*(3), 817–68. 1, 6

Fischbacher, U. (2007, #jun#). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178. 3.2

Gächter, S. and E. Fehr (2000, September). Cooperation and Punishment in Public Goods Experiments. *American Economic Review 90*(4), 980–994. 1

Gächter, S., E. Renner, and M. Sefton (2008). The long-run benefits of punishment. *Science 322*(5907), 1510–1510. 1

Gibbons, R. (1997). Incentives and careers in organizations. In D. M. Kreps and K. F. Wallis (Eds.), *Advances in Economics and Econometrics: Theory and Applications*, pp. 1–37. Cambridge, UK: Cambridge University Press. 17

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review 95*(1), 384–394. 3.5

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013). Measuring lying aversion. *Journal of Economic Behavior and Organization 93*, 293–300. 3.5

Grechenig, K., A. Nicklisch, and C. Thöni (2010). Punishment despite reasonable doubt—a public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies 7*(4), 847–867. 2

Greif, A. (1994, October). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy 102*(5), 912–950. 14, 5

Greiner, B. (2015, Jul). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125. 3.2

Guth, W., R. Schmittberger, and B. Schwarze (1982). An experimental-analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization 3*(4), 367–388. 1

Hall, B. (2000, December). Compensation and performance evaluation at arrow electronics. Technical Report 9-800-290, Harvard Business School. 18

Homans, G. C. (1961). *Social Behavior Its Elementary Forms*. New York, NY: Harcout, Brace & World, Inc. 15

Jensen, M. and W. Meckling (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics 3*(4), 305–60. 1

Kagel, J. H., C. Kim, and D. Moser (1996). Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior 13*(1), 100–110. 1

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986). Fairness and the assumptions of economics. *Journal of business*, S285–S300. 1

Kerr, S. (1975, December). On the folly of rewarding A, while hoping for B. *Academy of Management Journal 18*(4), 769–783. 18

Koszegi, B. (2014, December). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118. 1

Kreps, D. M. (1990). Corporate culture and economic theory. In J. E. Alt and K. A. Shepsle (Eds.), *Perspectives on Positive Political Economy*, pp. 90–143. Cambridge, U.K.: Cambridge University Press. 14, 5

Lazear, E. P. (1986). Salaries and piece rates. *Journal of Business 59*, 405–431. 4, 2

Leibbrandt, A. and R. López-Pérez (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization 84*(3), 753 – 766. 1

Lundquist, T., T. Ellingsen, E. Gribbe, and M. Johannesson (2009). The aversion to lying. *Journal of Economic Behavior & Organization 70*(1–2), 81–92. 3.5

MacLeod, W. B. (2003, March). Optimal contracting with subjective evaluation. *American Economic Review 93*(1), 216–240. 2

MacLeod, W. B. and J. M. Malcomson (1989, March). Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica 57*(2), 447–480. 14

MacLeod, W. B. and J. M. Malcomson (1998, June). Motivation and markets. *American Economic Review 88*(3), 388–411. 3, 14

Ostrom, E., J. Walker, and R. Gardner (1992). Covenants with and without a sword: Self-governance is possible. *American political science Review 86*(2), 404–417. 1

Prasnikar, V. and A. E. Roth (1992). Considerations of fairness and strategy: Experimental data from sequential games. *The Quarterly Journal of Economics 107*(3), 865–888. 1

Schelling, T. C. (1980). *The Strategy of Conflict.* Cambridge, MA: Harvard University Press. 16

Selten, R. (1975). Re-Examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory 4*, 25–55. 4

Simon, H. A. (1982). *Models of Bounded Rationality.* Cambridge, Mass.: MIT Press. 5

Syverson, C. (2011). What determines productivity?. *Journal of Economic Literature 49*(2), 326–365. 5

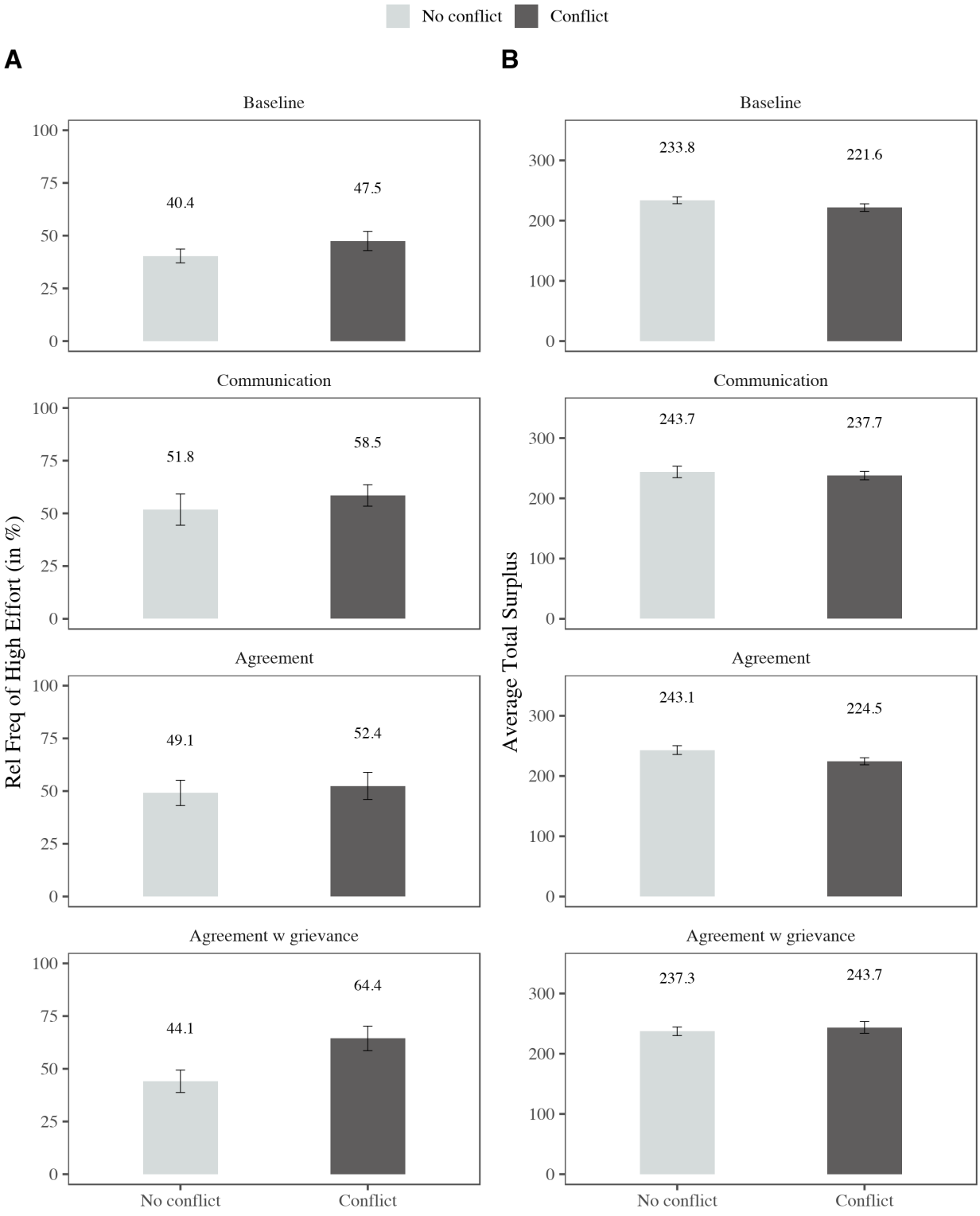
Tjosvold, D., A. S. Wong, and N. Y. Feng Chen (2014). Constructively managing conflicts in organizations. *Annu. Rev. Organ. Psychol. Organ. Behav. 1*(1), 545–568. 1

Townsend, R. (2007). *Up the Organization.* San Francisco, CA: John Wiley & Sons. (document)

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333. 4

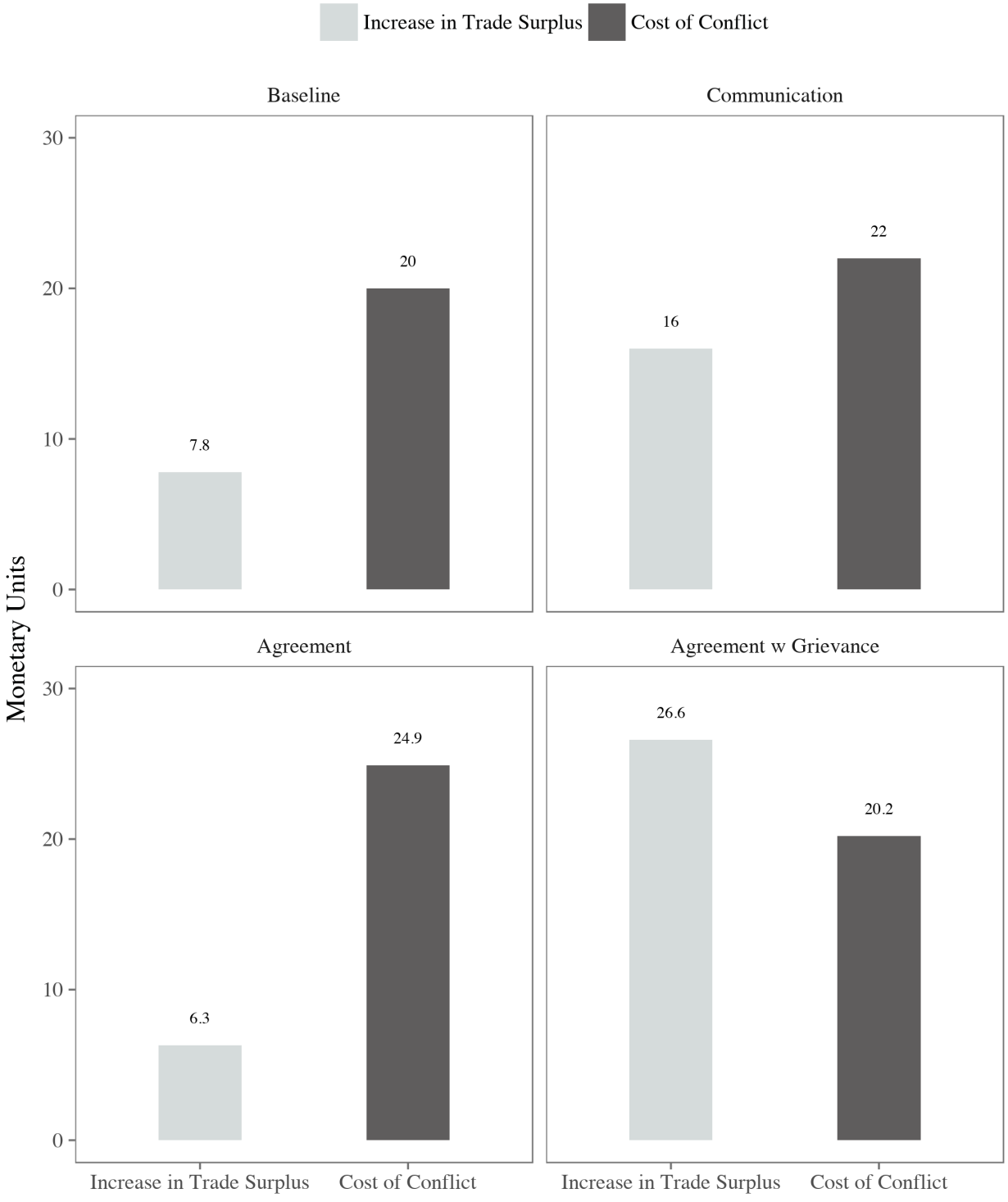
List of Figures

FIGURE 5.1. Effort and surplus



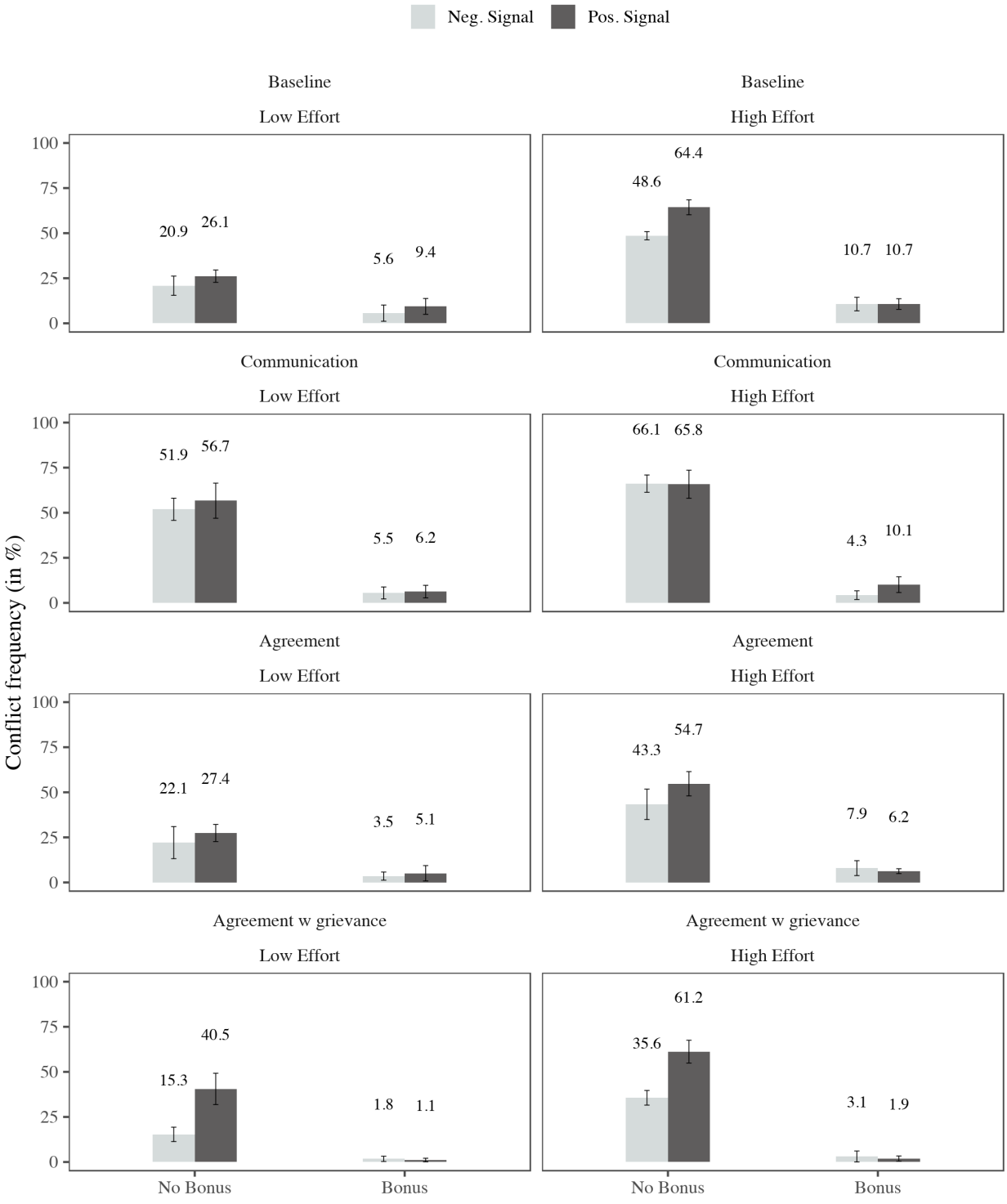
Note: Error bars represent plus/minus one standard error of the mean, clustered at the session level.

FIGURE 5.2. Gains and costs of conflict



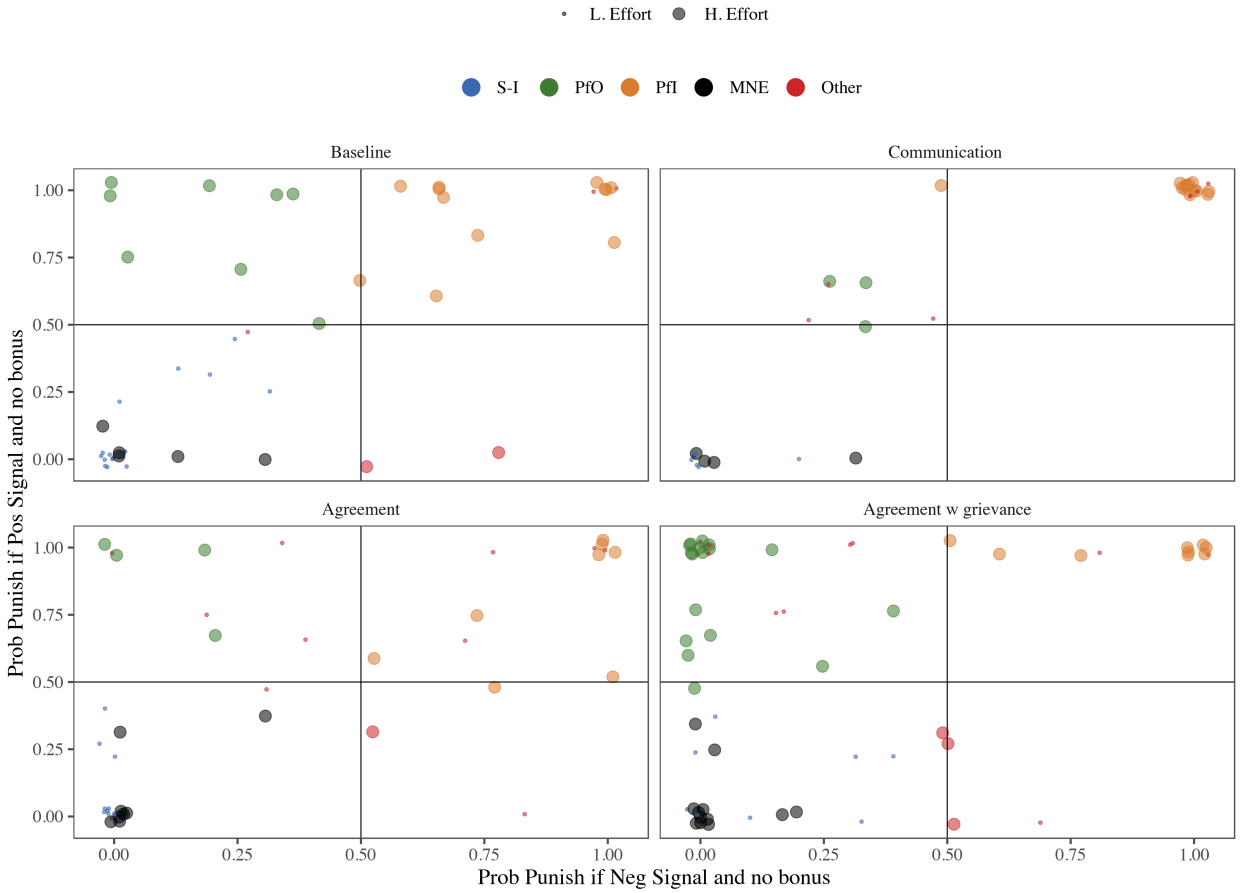
Note: Error bars represent plus/minus one standard error of the mean, clustered at the session level.

FIGURE 5.3. Determinants of conflict



Note: Error bars represent plus/minus one standard error of the mean, clustered at the session level.

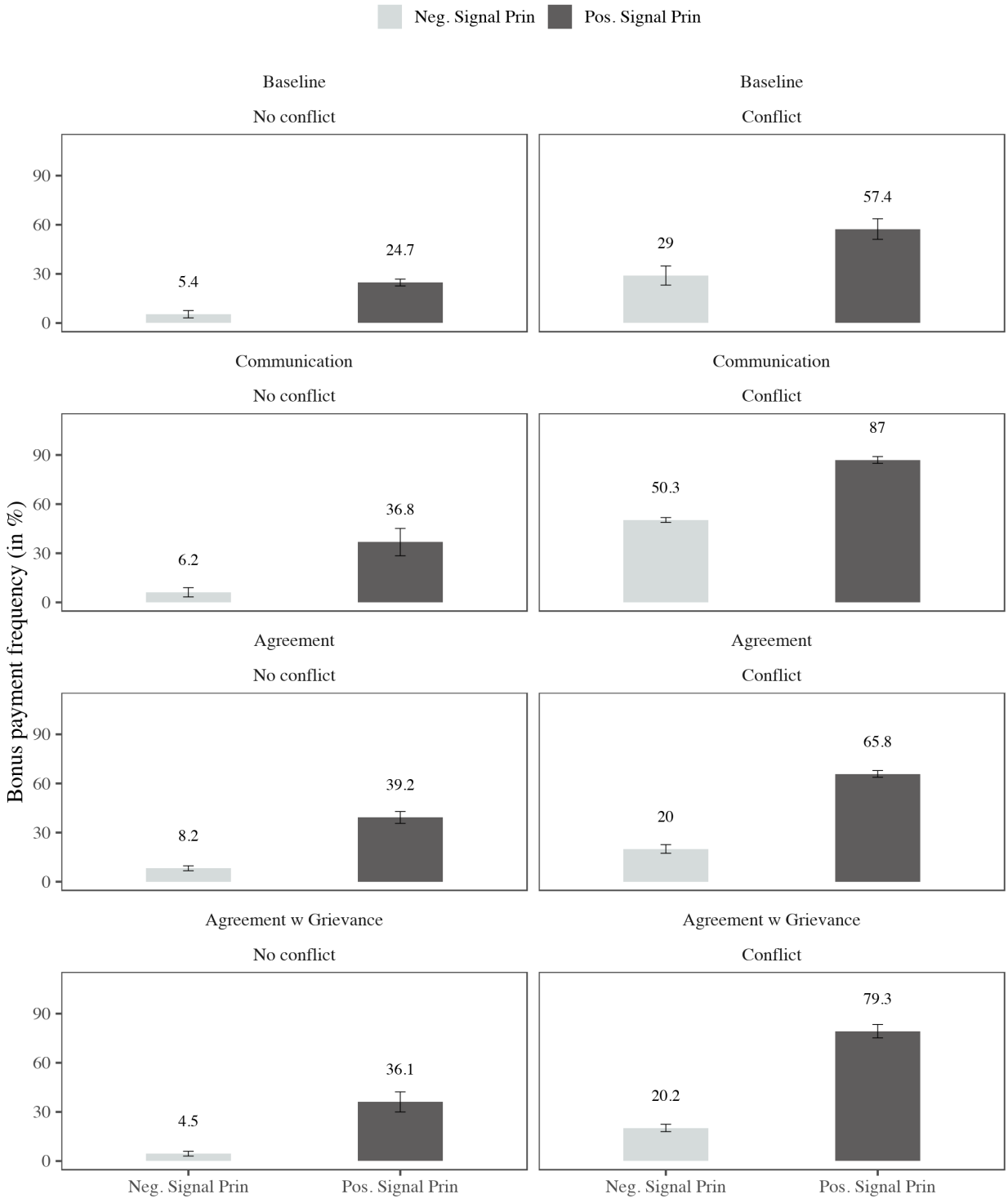
FIGURE 5.4. Worker's Effort and Conflict Decisions Averaged at the Individual Level



	S – I	PfO	PfI	MNE	Other
Baseline	36.2	17	25.5	10.6	10.6
Communication	19.4	8.3	36.1	11.1	25
Agreement	32.6	8.7	17.4	17.4	23.9
Agreement w grievance	14.8	27.9	14.8	19.7	23

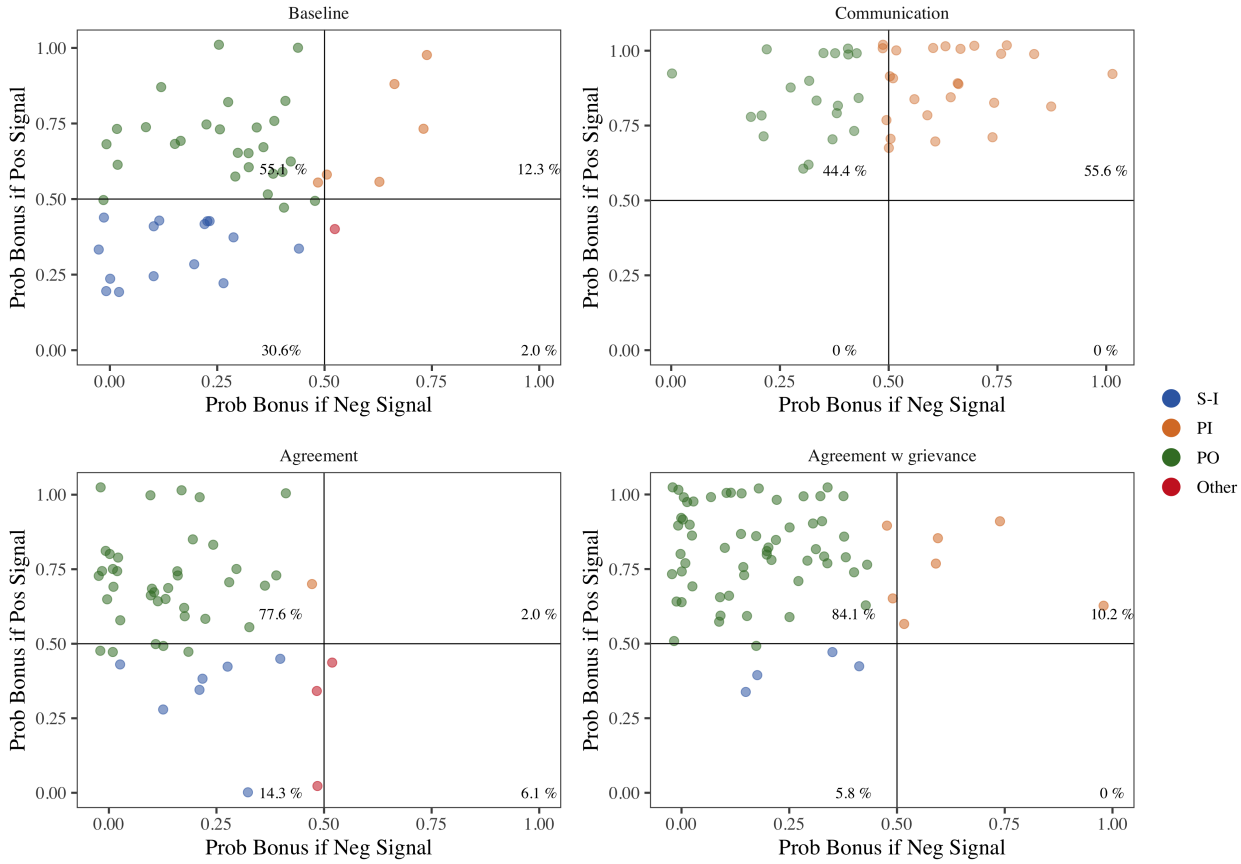
Note: In this graph each dot represents one worker in each treatment. The dots give information on (1) the level of effort, represented by size of the dot; and (2) the conflict behavior, represented by the position of the dot in the graph. The space of the graph is determined by the probability of conflict if the Worker does not receive a bonus and the signal is positive (y-axis) and the probability of conflict if the Worker does not receive a bonus and the signal is negative (x-axis). Furthermore, the colors of the dots help to emphasize the norm behavior that each worker most closely follows.

FIGURE 5.5. Determinants of bonus



Note: Error bars represent plus/minus one standard error of the mean, clustered at the session level.

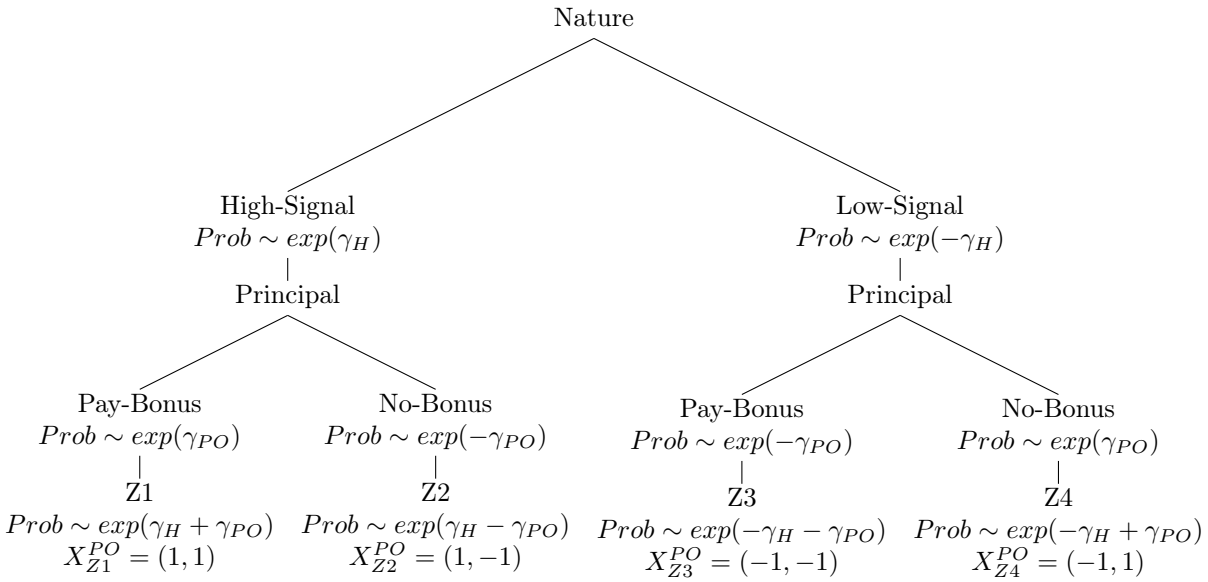
FIGURE 5.6. Principal's Bonus Payment Decisions Averaged at the Individual level



:

Note: In this graph each dot represents one individual in each treatment. The dots give information on the bonus payment behavior of the Principal. The space of the graph is determined by the probability of paying a bonus after a positive signal (y-axis) and the probability of paying a bonus after a negative signal (x-axis). Furthermore, the colors of the dots help to emphasize the norm behavior that each principal most closely follows.

FIGURE 5.7. Parameterizing the Principal's Choices under *Pay-for-Output*



List of Tables

TABLE 1. Social Surplus in Conflict Treatment

		Worker		
		Self-Interest (E=0)	Pay for Input (E=1)	Pay for Output (E=1)
Principal	Self-Interest	180	200	236
	Pay for Input	180	310	290
	Pay for Output	180	274	290

TABLE 2. Gain to Deviation from Norm in Conflict Treatment

		Worker					
		Self-Interest (E=0)		Pay for Input (E=1)		Pay for Output (E=1)	
		Principal	Worker	Principal	Worker	Principal	Worker
Principal	Self-Interest	0	0	-50	-20	-15	-16.75
	Pay for Input	-50	10	0	0	-16.25	-1.88
	Pay for Output	-16.25	-5.63	-16.25	-1.38	0	0

TABLE 3. Baseline - Development of effort and surplus over time

	Effort			Surplus		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Conflict	53.06	46.53	42.86	235.31	217.74	211.65
No conflict	52.83	34.34	33.96	259.43	216.76	225.10
p-values (RS)	0.753	0.151	0.421	0.222	1.000	0.310

TABLE 4. Baseline - Development of conflict rate over time (only if no bonus received)

	High Effort			Low Effort		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Positive Signal	59.52	59.46	81.81	18.52	26.92	31.43
Negative Signal	47.83	52.17	45.83	19.15	25.93	17.74
p-values (SR)	0.20	0.79	0.06	0.50	0.44	0.64

TABLE 5. Baseline - Development of bonus payments over time

	High Signal			Low Signal		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Conflict	54.20	60.18	58.33	30.70	28.03	28.47
No conflict	32.14	23.90	16.81	4.80	6.41	4.80
p-values (RS)	0.02	0.01	0.01	0.04	0.03	0.04

TABLE 6. Communication - Development of effort and surplus over time

	Effort			Surplus		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Conflict	60.89	45.58	57.78	243.28	234.71	235.12
No conflict	63.72	46.05	33.96	256.19	238.42	236.61
p-values (RS)	1.00	0.55	0.42	0.70	1.00	1.00

TABLE 7. Communication - Development of conflict rate over time (only if no bonus received)

	High Effort			Low Effort		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Positive Signal	63.64	62.50	72.73	71.43	50.00	37.50
Negative Signal	58.33	68.42	73.68	40.00	61.54	53.57
p-values (SR)	0.89	0.85	NA	0.10	NA	0.63

TABLE 8. Communication - Development of bonus payments over time

	High Signal			Low Signal		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Conflict	86.21	85.60	89.42	46.79	49.00	54.55
No Conflict	41.46	41.96	7.77	10.87	25.49	0.89
p-values (RS)	0.02	0.01	0.01	0.01	0.01	0.01

TABLE 9. Formal agreement with grievance - Development of effort and surplus over time

	Effort			Surplus		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Conflict	71.10	60.86	61.32	253.50	237.26	240.54
No Conflict	57.10	39.51	35.60	252.93	237.41	221.36
p-values (RS)	0.180	0.040	0.020	1.00	0.820	0.180

TABLE 10. Grievance - Development of conflict rate over time (only if no bonus received)

	High Effort			Low Effort		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Positive Signal	66.67	60.00	57.50	30.30	32.08	48.78
Negative Signal	40.00	44.44	37.04	22.50	13.73	11.86
p-values (SR)	0.040	0.040	0.460	0.860	0.160	0.320

TABLE 11. Formal agreement with grievance- Development of bonus payments over time

	Positive Signal			Negative Signal		
Periods	1-5	6-10	11-15	1-5	6-10	11-15
Conflict	81.58	78.38	77.90	18.59	20.61	21.38
No Conflict	42.26	37.95	27.03	7.05	4.43	2.29
p-values (RS)	0.00	0.00	0.00	0.01	0.03	0.00

TABLE 12. Empirical Norm Model for Principals

$Z_{Principal}$	Principal Signal	Bonus Paid?	X_Z^{PO}	X_Z^{PI}	X_Z^R
Z1	High	Yes	$(1,1)$	$(1,1)$	$(1,-1)$
Z2	High	No	$(1,-1)$	$(1,-1)$	$(1,1)$
Z3	Low	Yes	$(-1,-1)$	$(-1,1)$	$(-1,-1)$
Z4	Low	No	$(-1,1)$	$(-1,-1)$	$(-1,1)$

TABLE 13. Estimated Adherence of Principal's Behavior to a Norm

Experiment	**No Conflict Treatments**		**Conflict Treatments**	
	Pay for Output (γ^*_{PO})	Pay for Input (γ^*_{PI})	Pay for Output (γ^*_{PO})	Pay for Input (γ^*_{PI})
baseline	0.250^{***}	-0.894^{***}	0.298^{***}	-0.149^{***}
	(0.037)	(0.051)	(0.039)	(0.037)
Communication	0.288^{***}	-0.628^{***}	0.394^{***}	0.401^{***}
	(0.041)	(0.047)	(0.042)	(0.042)
Code of Conduct	0.330^{***}	-0.590^{***}	0.496^{***}	-0.145^{***}
	(0.039)	(0.043)	(0.042)	(0.038)
Code+Grievance	0.332^{***}	-0.687^{***}	0.678^{***}	0.043
	(0.034)	(0.040)	(0.038)	(0.031)

Note: *p<0.1; **p<0.05; ***p<0.01, errors clustered by session.

TABLE 14. Comparing Norms in Conflict Treatments

	Log Likelihood		Is a Norm Dominant at 1% significance levels		
Experiment	PO	PI/SI	PO	PI	SI
Baseline	**-867.7**	-1,010.16	**Yes**	No	No
(p-value)			(0.004)	(0.906)	(0.906)
Communication	-887	**-885.4**	No	No	No
(p-value)			(0.544)	(0.456)	(0.456)
Code of Conduct	**-916.6**	-986.5	**Yes**	No	No
(p-value)			(0.000)	(1.000)	(1.000)
Code+Grievance	**-1,250.66**	-1,444.25	**Yes**	No	No
(p-value)			(0.000)	(1.000)	(1.000)

Notes: P-values based upon non-nested Vuong where the null hypothesis is that the two norms, PO and PI are indistinguishable. The R norm is the same model as the PI norm, but with a change in the sign of the coefficient.

TABLE 15. Empirical Norm Model for Workers

Z_{Worker}	Worker Signal	Bonus Paid?	Principal Punished?	X_Z^{PO}	X_Z^{PI}	X_Z^R
Z1	High	Yes	No	$(1, 0, 1, 1)$	$(1, 0, 1, 1)$	$(1, 0, 1, 1)$
Z2	High	No	No	$(1, 0, -1, -1)$	$(1, 0, -1, -1)$	$(1, 0, -1, 1)$
Z3	Low	Yes	No	$(-1, 1, 0, 1)$	$(-1, 1, 0, 1)$	$(-1, 1, 0, 1)$
Z4	Low	No	No	$(-1, -1, 0, 1)$	$(-1, -1, 0, -1)$	$(-1, -1, 0, 1)$
Z5	High	Yes	Yes	$(1, 0, 1, -1)$	$(1, 0, 1, -1)$	$(1, 0, 1, -1)$
Z6	High	No	Yes	$(1, 0, -1, 1)$	$(1, 0, -1, 1)$	$(1, 0, -1, -1)$
Z7	Low	Yes	Yes	$(-1, 1, 0, -1)$	$(-1, 1, 0, -1)$	$(-1, 1, 0, -1)$
Z8	Low	No	Yes	$(-1, -1, 0, -1)$	$(-1, -1, 0, 1)$	$(-1, -1, 0, -1)$

TABLE 16. Estimated Adherence of Behavior to a Worker's Norm

Experiment	**Low Effort**			**High Effort**		
	(Self-Interest)	(Pay for Input)	(Pay for Output)	(Self-Interest)	(Pay for Input)	(Pay for Output)
Baseline	0.780^{***}	-0.057	0.466^{***}	0.336^{***}	0.521^{***}	0.536^{***}
	(0.067)	(0.051)	(0.057)	(0.057)	(0.061)	(0.061)
Communication	0.569^{***}	0.639^{***}	0.608^{***}	0.617^{***}	0.890^{***}	0.709^{***}
	(0.070)	(0.072)	(0.071)	(0.060)	(0.072)	(0.064)
Code of Conduct	0.793^{***}	-0.038	0.543^{***}	0.481^{***}	0.461^{***}	0.542^{***}
	(0.072)	(0.054)	(0.062)	(0.058)	(0.057)	(0.059)
Code+Grievance	0.876^{***}	0.070	0.740^{***}	0.682^{***}	0.595^{***}	0.811^{***}
	(0.073)	(0.052)	(0.067)	(0.048)	(0.046)	(0.052)
Observations	1,588	1,588	1,588	2,319	2,319	2,319
Log Likelihood	$-3{,}864.113$	$-4{,}150.913$	$-3{,}971.681$	$-5{,}837.546$	$-5{,}790.553$	$-5{,}691.499$
Note:						*p<0.1; **p<0.05; ***p<0.01, clustered by session

TABLE 17. Comparing Norms in Conflict Treatments

	Log Likelihood			Is a Norm Dominant at 1% significance levels		
	R	PI	PO	R	PI	PO
Baseline	-949.0	-925.9	**-923.8**	No	No	No
(p-value)				(0.998)	(0.595)	(0.405)
Communication	-971.3	**-923.5**	-955.4	No	**Yes**	No
(p-value)				(0.999)	(0.005)	(0.995)
Code of Conduct	-998.7	-1001.6	**-989.5**	No	No	No
				(0.822)	(0.897)	(0.178)
Grievance	-1,703.5	-1,729.1	**-1,664.8**	No	No	**Yes**
				(0.992)	(1.000)	(0.008)

Notes: P-values based upon non-nested Vuong test where the null hypothesis is that two norms are indistinguishable compared to a single norm. The comparison is always made to the best alternative model as measured by the likelihood value.

TABLE 18. Estimated Adherence of Behavior to a Work Culture

Experiment	Cultural Fits with Conflict		
	(Self-Interest)	(Pay for Input)	(Pay for Output)
Baseline	0.462***	0.391***	0.427***
	(0.059)	(0.047)	(0.047)
Communication	-0.265**	0.845**	0.698***
	(0.074)	(0.044)	(0.044)
Code of Conduct	0.438***	0.458***	0.578***
	(0.047)	(0.046)	(0.045)
Code+Grievance	0.240***	0.711***	0.889***
	(0.044)	(0.034)	(0.034)

Note: *p<0.1; **p<0.05; ***p<0.01

TABLE 19. Comparing Cultures in Conflict Treatments

	Log Likelihood			Is a Norm Dominant at 1% significance levels		
Experiment	SI	PI	PO	SI	PI	PO
Baseline	**-2,346.5**	-2,364.8	-2,358.2	Yes	No	No
(p-value)				(0.1)	(0.90)	(0.78)
Communication	-2,186.4	**-2,023.7**	-2,082.4	No	Yes	No
(p-value)				(1.00)	(0.00)	(1.00)
Code of Conduct	-2,264.0	-2,259.4	**-2,229.2**	No	No	Yes
				(0.98)	(1.00)	(0.02)
Code+Grievance	-3,384.3	-3,208.1	**-3,086.5**	No	No	Yes
				(1.00)	(1.00)	(0.00)

Notes: P-values based upon non-nested Vuong test where the null hypothesis is that two norms are indistinguishable compared to a single norm. The comparison is always made to the best alternative model as measured by the likelihood value or p-value.