Dhami, Sanjit; Wei, Mengxing

Working Paper

The Incentive Compatibility Condition, Firm Culture, and Social Norms under Moral Hazard: Theory and Evidence

CESifo Working Paper, No. 11371

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Dhami, Sanjit; Wei, Mengxing (2024) : The Incentive Compatibility Condition, Firm Culture, and Social Norms under Moral Hazard: Theory and Evidence, CESifo Working Paper, No. 11371, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/305613

CESifo WORKING PAPERS

11371
2024

September 2024

The Incentive Compatibility Condition, Firm Culture, and Social Norms under Moral Hazard: Theory and Evidence

Sanjit Dhami, Mengxing Wei

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11371*

The Incentive Compatibility Condition, Firm Culture, and Social Norms under Moral Hazard: Theory and Evidence

Abstract

In a principal-agent model under moral hazard we examine the psychological and social motivations of the agent that influence the incentive compatibility condition (ICC) of the agent. Under "firm culture" firms emphasize that high effort is consistent with its culture. Under "industry-wide social norms" external to the firm, the social group emphasizes high effort levels. We only consider the case where the ICC is violated in the classical case. A significant fraction of the agents choose high effort. Firm culture backed by simple disapproval of low effort is more effective relative to our baseline under fixed wages. Strong social norms are as effective as firm culture under variable wages, but more effective under fixed wages. Firm culture dominates weak social norms. Variable wages induce high effort (incentive effects) but also crowd out intrinsic motivation in the form of (i) guilt aversion from not following firm culture and (ii) shame aversion from not following social norms.

JEL-Codes: D010, D910.

Keywords: incentive compatibility, insurance and incentives, firm culture, guilt-aversion, social norms, shame-aversion.

Sanjit Dhami
Department of Economics, School of Business
University of Leicester / United Kingdom
sd106@le.ac.uk

*Mengxing Wei**
School of Economics, Laboratory for Economic Behaviors and Policy Simulation
Nankai University, Tianjin / China
mengxing.wei@hotmail.com

*corresponding author

September 12, 2024

1 Introduction

The *incentive compatibility condition* (ICC) is critical to many areas of economics. The classical ICC in principal-agent problems typically assumes that workers are only motivated by monetary incentives. Throughout, we are only interested in situations where the ICC is violated in the classical principal-agent model, under fixed and variable wages. Hence, the classical model predicts that workers will choose low effort. We show that "firm culture" and "social norms" might nevertheless induce workers to choose a high effort level. We also consider the crowding out of intrinsic motivation by monetary incentives such as variable wages, when fixed wages are available. We only consider classical wage-based incentives, and abstract from issues of the optimality of contracts, or other forms of incentives such as bonuses and trust contracts (Fehr and Falk, 2002; Fehr and List, 2004; Fehr et al., 2007).

1.1 Classical principal-agent problem

We consider a simple principal-agent model under moral hazard in which a worker (agent) has one of two possible effort choices, a high effort and a low effort. Higher effort is relatively more costly. The level of effort translates probabilistically into one of two states– high output (good state) or a low output (bad state) for the firm. Higher effort makes the good state more likely relative to low effort. The firm finds it more profitable to induce higher effort from the worker, but the effort is either (i) unobserved to the firm and/or (ii) not verifiable to a third party (moral hazard). The classical solution is to offer state-contingent wages to the worker such that the worker voluntarily prefers to exert high effort. This is encapsulated in the *incentive compatibility condition* (Mirrlees, 1971; Holmstrom, 1979). In addition, the *individual rationality condition*, IRC, of the worker must be satisfied.

The key tradeoff in the classical analysis is between providing insurance and incentives to the worker. In general, the standard model predicts optimal incentive schemes that are too high-powered (highly variable wages) relative to the relatively flat wages observed in the real world. Holmström and Milgrom (1991, p. 24) write: "...it remains a puzzle for this theory that employment contracts so often specify fixed wages and more generally that incentives within firms appear to be so muted, especially compared to those of the market." One potential explanation is that workers might also be motivated by non-monetary incentives such as firm culture and social norms that loosen the ICC, requiring lower-powered incentives.[1]

1.2 Firm culture

The term "firm culture" comprises the set of informal rules or institutions that are used to guide the actions, beliefs (Bénabou, 2013) of employees, their commitment problems (Greif, 1994), and reputations (Tirole, 1996). Some have characterized firm culture as norms within organizations or internal norms (Akerlof, 1982; Kreps, 1990; Huck et al., 2012; Chatman and O'Reilly, 2016); conflicts of interests within firms (Cyert and March, 1963); or culture as one of

[1] Other-regarding preferences is yet another possibility (Fehr and Falk, 2002; Fehr and List, 2004; Fehr et al., 2007) but our design ensures that other-regarding considerations do not arise.

the productive assets of a firm in the form of organizational capital (Dessein and Prat, 2022). In a study of 1348 North American executives, 91% of executives considered corporate culture to be "important" or "very important" at their firm, and 92% believed that culture improves firm value (Graham et al., 2022).[2]

Firm culture can, and does, in real life influence a range of key aspects of a firm's operation. However, we define firm culture in the limited sense of expectations of the firm (principal) of the effort levels from the worker (agent) that are consistent with the firm culture. Indeed, the very first of 12 question that is asked in Gallup polls on "employee engagement" in firms is to comment on the following: *I know what is expected of me at work.*

If the firm culture emphasizes expectations of a high effort level, then the worker might feel "guilty" from falling behind those expectations, even if the ICC, derived from purely monetary considerations, is violated. The literature on psychological game theory has formalized guilt-aversion (Battigalli and Dufwenberg, 2007). Guilt aversion plays a critical role in motivating worker effort, even when the false consensus effect on belief elicitation is accounted for (Khalmetski et al., 2015; Dhami et al., 2019; Dhami et al., 2022; Dhami et al., 2023).[3]

Kandel and Lazear (1992) distinguish between "internal pressure" that arises from guilt-aversion by, for instance, not following firm culture and "external pressure" that arises from shame aversion by, for instance, not following social norms outside the relationship. Unlike shame that is triggered from violating social norms (see section 1.3 below), guilt is typically triggered in close communal relationships in which one believes one has caused harm, loss, or distress to a relationship partner (Baumeister et al., 1994). We follow this distinction between guilt that arises from violating firm culture and shame that arises from violating social norms.

Norms, that are "internal" to the firm (which we treat under the heading "firm culture") are often conveyed to the workers in the form of *performance standards*. By contrast, norms that are "external" to the firm (which we term "social norms" or "industry wide standards"; see Section 1.3 below) are conveyed through informal social networks or direct personal observation of the behavior of one's social group.

1.3 Social norms

Akerlof (1982) identified the effects of norms on the effort level of workers more generally by allowing for norms that lie both inside the firm (firm culture) and outside the firm (industry-wide norms). There are the following two important reasons to study norms arising from outside the principal-agent relationship more seriously; this also ties in with earlier research on norms within organizations (Kreps, 1990; Huck et al., 2012; Chatman and O'Reilly, 2016). (i) In real world organizations, there are multiple divisions, and within each division there are smaller

[2] For a survey of the cross-country cultural differences in organizations, see Hofstede et al. (2010). Culture might also involve the decision on how much power to decentralize to different levels (Besley and Persson, 2022) and how identities shape cultural values within firms (Akerlof and Kranton, 2000; Besley and Ghatak, 2005). The other social sciences have also tried to formalize firm culture and its affect on firm performance in a variety of ways (Whyte, 1956; Hofstede, 1984; Wilson, 1989; Schein, 1990).

[3] "Direct belief elicitation" which involves directly asking players about their second order beliefs (critical in the formation of guilt-aversion) is subject to the false consensus effect and produces unreliable estimates and effects of guilt-aversion (Ellingsen et al., 2010). A rich literature demonstrates the importance of guilt-aversion under direct belief elicitation; see Battigalli and Dufwenberg (2022) for a survey.

micro entities in the form of multiple subdivisions (or even sub-subdivisions). The principal-agent problem also gets played out in these smaller micro entities. The overarching values, culture and norms of the firm are in some sense external to the specific principal-agent relation within the micro entity. (ii) Workers do not operate in isolation from the rest of society of which they are a part, nor in isolation from other firms within the same industry.

We model social norms using the framework in Bicchieri (2006) and Elster (2011) and formalized further and applied, for instance, in Dhami et al. (2022) and Dhami et al. (2023). Social norms have three components. *Empirical expectations* (expectations of the effort undertaken by others in the social group); *normative expectations* (the worker's beliefs about the effort that the relevant social group expects that others in the group 'ought' to exert), and (3) *Sanctions* by members of the social group for violating the social norms for effort.[4] Adherence to normative standards of behavior incorporated in social norms is underpinned by the emotion of "shame" from violating social norms (Bicchieri, 2006; Elster, 2011; Gintis, 2017).

Which of the two emotions, guilt or shame, elicit higher worker effort? Cultural anthropologists sometimes distinguish between guilt cultures (typically Judeo-Christian religions) and shame cultures (typically Arabic cultures and Eastern religions). Kitayama et al. (2006) distinguish between Japanese and US subjects on these basis and Bedford (2004) traces the acceptance of shame in China to the teachings of Confucius. Dhami et al. (2022) show from Pakistani data that effort choice in microfinance contracts is dictated by shame-aversion rather than guilt-aversion, when both emotions are operative.

1.4 Treatments

We first outline our theory in the form of a simple principal-agent model which we faithfully implement in pre-registered experiments. We consider three main treatments in our model: Baseline treatment T0, i.e., the classical principal-agent problem (Section 1.1, above); treatment T1, i.e., firm culture supported by guilt-aversion (Section 1.2, above); treatment T2, i.e., social norms supported by shame-aversion (Section 1.3). We follow a between-subjects design.

For our chosen parameter values, the ICC is always violated in treatment T0, so if the classical model captured all relevant human motivations we should never observe high effort in our data. Yet, humans might also be motivated by internal moral norms of behavior (intrinsic motivation), even in treatment T0, and might experience guilt from the violation of internal moral rules and imperatives (Freud, 1930/1961; Lazarus, 1991; Gintis, 2017). Relative to the baseline treatment, the presence of firm culture (T1) and social norms (T2) are predicted to increase average effort as the worker experiences, respectively, guilt aversion and shame aversion from violating internal firm norms and external industry-wide norms.

We split treatment T1 (firm culture) into two sub-treatments: T1N (firm cannot sanction workers for effort because effort is unobservable) and T1S (firm can impose non-monetary sanctions such as disapproval on workers based on observable but non-verifiable effort). In static classical principal-agent problems, both methods of implementing moral hazard (unobservable

[4] In our experiments, the social group can impose non-monetary sanctions, such as disapproval, on workers who violate a social norm of high effort. It is well known that non-monetary sanctions, such as disapproval, are reasonably effective and often as effective as monetary sanctions (Masclet et al., 2003).

effort and observable but unverifiable effort) are predicted to produce identical outcomes. Due to potentially heightened guilt-aversion in T1S, we expect higher average effort in T1S relative to T1N and T0.

In treatment T2 (social norms), we inform subjects, based on a previous pilot of the same experiment, that $x\%$ of their social group has the normative expectation that they should exert a high effort level; we also ensure satisfaction of the other conditions for a social norm, noted above. In other words this is a public signal of the normative expectations of the social group (normative injunction). This splits treatment T2 into two sub-treatments: T2L (low value of $x\%$) and T2H (high value of $x\%$). A stronger normative injunction in treatment T2H creates even higher shame aversion from violating the social norm, so it is predicted to increase average effort relative to treatment T2L.

We examine the relative effects of variable wages and fixed wages within each treatment. Variable wages are known to loosen the ICC in the classical model through incentive effects, inducing higher effort. Empirical evidence suggests that when firms offer incentive contracts (variable wages) in the presence of alternatives that are low powered (fixed wage contracts) it is interpreted as "hostile intent" on the part of the firm by workers, which influences the optimal choice of effort (Fehr and Falk, 2002; Dhami, 2019. Vol. 2). As such, the influence of guilt-aversion and shame-aversion might vary under variable wages relative to fixed wages. Our theoretical model highlights the relevant tradeoffs and we test its predictions.

1.5 Related literature

A separate literature, not directly related to our paper, looks at the effects of other factors, which might also influence the ICC. This includes pre-play promises (Vanberg, 2008; Charness and Dufwenberg, 2006); behaviorally motivated contracts, such as trust contract and bonus contract in static and dynamic contexts (Fehr et al., 2007; Brown et al., 2004); and self selection of workers of different productivity into different contractual forms (Dohmen and Falk, 2011). We distinguish our paper from traditional models of gift exchange (Akerlof, 1982, Fehr et al. 1993) and from belief-based accounts of gift exchange (Dhami et al., 2023) by considering the role of firm culture and social norms for the ICC in a principal-agent problem under uncertainty.

1.6 Experiments and findings

We collected our data from lab experiments conducted in China with 415 students from Nankai University. Our findings are as follows. Effort is higher under firm culture with sanctions (treatment T1S) relative to the baseline treatment in T0. The percentage of high effort choices in the presence of the high signal of normative expectations (treatment T2H) is significantly higher than the treatment with low signal of normative expectations (treatment T2L) at all effort levels in both the variable and fixed wage cases. Furthermore, relative to the baseline treatment, T0, social norms enhance effort when the social norms are strong (T2H), but not when social norms are weak (T2L).[5] The observed treatment effects between T0 and T2 in

[5] We note that due to the constraints of selecting the normative injunction from an actual previous experiment, our "strong" norms are only "medium strong" and "weak" norms are "quite weak".

the fixed and variable wage cases confirm that subjects experience shame from violating social norms, and they are programmed to conform to normative injunctions (Bicchieri, 2006; Gintis, 2017).

What is the relative effectiveness of firm culture and social norms? The percentage of high effort choices is significantly greater in treament T2H relative to firm culture in the absence of sanctions (treament T1N); and for the fixed wage case, also greater than treatment T1S (firm culture in the presence of sanctions). Overall, the highest effort always arises in treatment T2H. In this sense, social norms backed by shame aversion achieve the highest effort in all cases and this fits in with the available lab experimental evidence (Dhami et al., 2022). However, under variable wages, the differences between T1S and T2H are not statistically different, even though the average effort under T2H is still higher. This suggests that when offering variable wages, firms can emphasize firm culture but when offering fixed wages, firms will do better by emphasizing social, industry-wide, norms.[6]

Despite the violation of the ICC, effort is higher under variable wages relative to fixed wages. This arises purely on account of incentive effects of wages and this result is also consistent with other literature where the ICC is not violated (Eriksson and Villeval, 2008; Lazear, 2000; Paarsch and Shearer, 2000). In our probit regressions, we observe higher treatment differences in high effort between T1S and T0 (and T1N and T0) for the fixed wage case but smaller differences for the variable wage case. Furthermore, the differences T1N–T0 are higher relative to T1S–T0. We hypothesize that two kinds of crowding-out effect of intrinsic motivation are likely to explain these result. (i) Guilt-aversion from violating firm culture that enhances high effort in treatment T1 applies to a greater extent under fixed wages relative to variable wages. This ties in with the literature that extrinsic motivation (in the variable wage case) can crowd out intrinsic motivation (in the fixed wage case), and guilt-aversion is an important form of intrinsic motivation (Gneezy and Rustichini, 2000; Bénabou and Tirole, 2003, 2006). (ii) Crowding-out of guilt aversion is likely to be stronger when the firm is perceived to take a hostile action, i.e., in the presence of sanctions by the firms on workers (T1S vs T0) as compared to the lack of sanctions (T1N vs T0); and this is what we find.

Our data is consistent with workers suffering from a "fixed" level of guilt and shame from violating, respectively, firm culture and social norms, rather than "variable" levels of guilt and shame arising from the extent of violation of firm culture and social norms.

The plan of the paper is as follows. Section 2 outlines the model. Section 3 derives the theoretical predictions. Section 4 describes the experimental design and Section 5 gives the experimental results. We conclude in Section 6. All proofs are in the appendix.

[6]Consider the following anecdotal evidence from a related domain. University salaries are typically closer to fixed wages than high powered variable wages. In terms of research and teaching evaluations, many university departments emphasize matching best practice in similarly ranked university departments in the country, which corresponds to industry-wide norms of effort in our model.

2 Model

Consider a static principal-agent model in the presence of moral hazard. We abstract from issues of optimal contract design and are interested in the behavior of workers, given standard state-dependent wage contracts chosen by a firm in classical principal-agent models. A firm (principal) hires a worker (agent) to work on a project. The worker chooses an effort level e_j, $j = L, H$ at a cost c_j such that $0 \leq e_L < e_H$ and $0 \leq c_L < c_H$. The worker's effort is unobservable to the firm, hence effort cannot be written as part of the contract (moral hazard).[7] The worker's outside option is a monetary amount $\overline{u} \equiv 0$.

The firm's production technology is stochastic, giving rise to two possible states of the world. A good state, $s = g$, in which the output is π_1 and a bad state, $s = b$, in which the output is π_0: $0 \leq \pi_0 < \pi_1$. The worker's effort level induces a conditional probability distribution over the two states. When effort is high, e_H, the probability of the good (bad) state is $p_H > 0$ $(1 - p_H)$. When effort is low, e_L, the probability of the good (bad) state is $p_L > 0$ $(1 - p_L)$ such that

$$p_L < p_H. \tag{2.1}$$

Thus, the high (low) output level is more likely when effort is high (low).

The state, $s = b, g$ is observed by both parties and verifiable to a third party, hence, it can be written down as part of the contract. The firm privately observes its state dependent output levels π_0, π_1, while the worker only knows that the firm's profits in the good state are higher. This rules out considerations of other-regarding preferences among workers in their decisions. Hence, the firm offers the following contract to the worker: If the state is good, $s = g$, the worker receives the wage w_1 and if the state is bad, $s = b$, the worker receives the wage w_0. The contract offered by the firm might also include an injunction for the worker to exert a particular level of effort, e_j, $j = L, H$, hence, the contract may be summarized by $\{w_0, w_1, e_j\}$.

In the classical principal-agent framework, the worker's expected utility from effort e_j is,

$$EU(e_j) \equiv U_j = (1 - p_j)u(w_0) + p_j u(w_1) - c_j; j = L, H. \tag{2.2}$$

The firm wishes to maximize expected profits, given by

$$E\pi(e_j) = E\pi_j = (1 - p_j)(\pi_0 - w_0) + p_j(\pi_1 - w_1); j = L, H. \tag{2.3}$$

To allow for more compact notation, we use the delta symbol, $\triangle$, for changes in various magnitudes– utility, profits, wages, costs, probabilities, effort levels– as follows.

$$\Delta u = u(w_1) - u(w_0);\ \Delta\pi = \pi_1 - \pi_0;\ \Delta w = w_1 - w_0;\ \Delta c = c_H - c_L;\ \Delta e = e_H - e_L;\ \Delta p = p_H - p_L. \tag{2.4}$$

We only consider the case where the firm always prefers the high effort level e_H, i.e., $E\pi_L < E\pi_H$. Using (2.3), (2.4) we get $E\pi_L < E\pi_H \Leftrightarrow \Delta w < \Delta\pi$, and we ensure that this condition is met in our experiments.

[7] This is true for all our treatments except T1S (described below) where the firm cannot produce a verifiable signal of effort to a third party, which is another, equivalent, way of implementing moral hazard in the classical principal-agent model.

3 Predictions of the theoretical model

In this section, we consider the predictions of our theory in three different main treatments, T0, T1, T2.

3.1 Treatment T0: The classical principal-agent analysis

Treatment 0 is our baseline treatment and the sequence of moves is as follows.

1. Stage 1: The firm announces a state-dependent contract $\{w_0, w_1\}$.
2. Stage 2: The worker chooses the optimal effort $e^* \in \{e_L, e_H\}$.
3. Stage 3: The realization of the state of the world, $s = b$ or $s = g$ is publicly revealed (profit levels in these states, π_0 or π_1, are privately observed by the firm). The expected utility of the worker is given by (2.2) and the expected profit of the firm is given by (2.3).

The ICC ensures that the worker prefers effort e_H to e_L; it requires $U_H > U_L$.[8] Using (2.2):

$$ICC : U_H > U_L \Leftrightarrow \triangle p \Delta u > \Delta c. \tag{3.1}$$

The individual rationality (IRC) constraint is satisfied if $U_H \geq \overline{u} \equiv 0$. Using (2.2):

$$IRC : U_H \geq \overline{u} \Leftrightarrow u(w_0) + p_H \Delta u - c_H \geq 0. \tag{3.2}$$

The IRC always holds in our experiments, so we do not discuss it further.

From (3.1), the ICC is violated if

$$\Delta u < \frac{\Delta c}{\triangle p}. \tag{3.3}$$

Suppose that subjects are approximately risk neutral over small stakes, hence, $\Delta u = u(w_1) - u(w_0) \approx w_1 - w_0$. For such subjects, the violation of the ICC in (3.3) can be rewritten as:

$$w_1 < w_0 + \frac{\Delta c}{\triangle p} \equiv a_1. \tag{3.4}$$

If the ICC is violated for a risk neutral individual, as above, it is also violated for a risk averse subject for any level of risk aversion. This is shown in Proposition 1; the simple proof is relegated to the appendix. Hence, our results also hold in the presence of risk aversion. This holds true in all our treatments.

Proposition 1. *Suppose that (3.4) holds, so that the ICC is violated for a risk neutral individual. Then, for any level of risk aversion, the ICC in (3.1) is also violated.*

The condition in (3.4) and Proposition 1 ensure that even if subjects in our experiments are risk averse for small stakes, the relevant ICC is violated for them. Thus, by choosing the contractual parameters such that the ICC will fail for risk neutral workers, we ensure that it fails for all workers, irrespective of their level of risk aversion.

We abstract from reputational and repeated-game concerns. While such concerns are undoubtedly important in real-world employment relationships, we wish to discover if other, non-reputational, mechanisms can also induce high effort even when the ICC is violated, as in (3.3).

[8] Throughout we use the tie-breaking rule that when workers are indifferent between the two actions, $U_H = U_L$, they chose low effort. Nothing of significance for our results hinges on this particular rule.

Remark 1. *(Internal moral norms) Unlike treatments T1 and T2, treatment T0 lacks any notion of firm culture and social norms of efforts that create additional psychological and social inducements for workers to exert high effort. However, this does not imply that workers in treatment T0 are 'amoral'. In particular, workers might have their own internal moral norms of high effort, say, based on reciprocity and gift exchange, so they might suffer from "guilt-aversion" from exerting low effort relative to their internal moral norms in treatment T0 (Freud, 1930/1961; Lazarus, 1991; Elster, 2011; Gintis, 2017). Such workers might wish to exert high effort even if the ICC is violated in the classical principal-agent model and internal moral norms might apply to all treatments. Subsection 3.5 below expands further on this point, draws on issue of intrinsic vs extrinsic motivation, and briefly sketches how this might be modelled.*

3.2 Treatments T1N and T1S: Guilt aversion in principal-agent contracts

Treatment 1 considers the effects of firm culture, or internal corporate norms, that are mediated through the channel of the worker's guilt-aversion from falling below the expectations of firm culture. Within T1, we have two sub-treatments, T1N and T1S. In treatment T1N (firm culture with "no sanctions") the firm does not observe the worker's effort level (non-observable effort). In treatment T1S (firm culture with "sanctions"), the firm can observe the effort of the worker, but cannot produce verifiable evidence of low effort to a third party in order to impose monetary sanctions on the worker (observable but non-verifiable effort). In the classical static principal-agent model, both methods of implementing moral hazard produce identical outcomes. However, in treatment T1S, the firm can impose non-monetary sanctions on the worker for low effort, such as disapproval, which distinguishes it from treatment T1N.[9] It has been shown that social disapproval is, by itself, a powerful and effective mechanism that influences behavior (Bowles and Gintis, 2011; Dhami, 2019, Vol. II). A related advantage of treatment T1S is that it allows us to directly compare the effects if firm culture with social norms in treatment T2, where the social group can impose identical non-monetary sanctions on the worker for exerting low effort (Section 3.3 below).

In treatment T1N (no sanctions), the sequence of moves is identical to T0 in Section 3.1 except that a new stage, Stage 0, precedes Stage 1. Stages 2, 3 are as in T0.

Stage 0 in T1N and T1S: The firm announces a "mission statement" outlining the "firm culture" or internal social norms of the firm. This statement emphasizes that the high effort level e_H is consistent with the firm culture, but e_L is not.

Treatment (T1S) is identical to treatment T1N, except that the firm can impose non-monetary sanctions on the worker in a new stage, Stage 4, that follows Stage 3.

Stage 4 in T1S: The firm can express non-monetary disapproval of the effort choices of the workers.[10]

In the presence of firm culture and internal firm norms of high effort, e_H, a guilt-averse worker might feel guilty from choosing effort below the expectations of the firm. We use the term guilt-aversion in the sense in which Battigalli and Dufwenberg (2007) use "simple guilt."

[9] Among all our treatments, only in treatment T1S is the firm able to observe the effort level of the worker.

[10] In our experiments, disapproval takes the form of showing a red thumbs-down sign on a computer screen to the worker for 3 minutes.

Simple guilt arises when player A chooses an action that is below player A's beliefs about the action that another player, player B, expects from player A. The Stage 0 announcement of an effort level of e_H in treatments T1N and T1S makes explicit the expectations (first order beliefs) of the firm to the worker so the worker can accurately form beliefs about the expectations of the firm (second order beliefs of the worker). In this case, the worker's second order beliefs assigns a probability 1 that the firm expects it to undertake the effort level e_H.

The worker suffers a guilt cost that is non-decreasing in the shortfall in effort relative to the firm's expectations. This is captured by a guilt aversion function $g : \mathbb{R} \to \mathbb{R}$, which captures the guilt from choosing an effort e_L when the firm expects an effort level e_H from the worker.[11]

$$g(e_j) = \phi\left(e_H - e_j\right), \, j = L, H, \tag{3.5}$$

where ϕ is an non-decreasing function, $\phi' \geq 0$, such that $\phi(0) = 0$. Thus, the worker suffers no guilt when choosing an effort level $e_j = e_H$. For example, $\phi = a + b\sqrt{x}$, $a, b > 0$, is a concave guilt aversion function which captures diminishing marginal guilt sensitivity. The case $a > 0$, $b = 0$ corresponds to a constant guilt-aversion function where the worker suffers a fixed amount of guilt, irrespective of the amount by which the worker fails to meet the firm's expectations. This corresponds to the case $\phi' = 0$ and turns out to be particularly empirically relevant for us.

In the presence of guilt-aversion, the worker's utility function from choosing the effort level e_j, is given by

$$V(e_j) \equiv V_j = [(1 - p_j)u(w_0) + p_j u(w_1) - c_j] - \lambda_k \phi\left(e_H - e_j\right); \, j = L, H, \, k \in \{T1N, T1S\}. \tag{3.6}$$

where $\lambda_k \geq 0$ is the relative weight placed on guilt aversion and $k \in \{T1N, T1S\}$ is a treatment index that captures the effects of non-monetary sanctions such as disapproval (treatment T1S) or their absence (treatment T1N). We assume that

$$0 \leq \lambda_{T1N} < \lambda_{T1S}, \tag{3.7}$$

i.e., guilt aversion bites more when the worker faces sanctions from the firm, relative to the case when there are no sanctions. Sanctions, because they involve disapproval by the other player, heighten the effects of guilt-aversion. Comparing the objective functions in T0 and T1, (2.2) and (3.6), the only difference is the last term in (3.6) that captures the disutility from guilt aversion. For a worker who suffers no guilt aversion, i.e., $\lambda_k = 0$, the objective functions in T0 and T1 are identical ($V_L \equiv U_L$). We allow for heterogeneity in preferences among workers, but avoid further notation for this purpose. Since the worker faces no guilt from choosing the effort level e_H, we have $V_H \equiv U_H$, where U_H is defined in (2.2).

From (3.6), the ICC in the presence of guilt aversion ensures the worker will choose e_H:

$$ICC : V_H > V_L \Leftrightarrow (\triangle p)\, \Delta u > \Delta c - \lambda_k \phi\left(\Delta e\right); \, k \in \{T1, T2\}. \tag{3.8}$$

In the absence of guilt aversion, the last term on the RHS in (3.8) equals zero. Hence, ceteris-paribus (i) the ICC is easier to satisfy in the presence of guilt aversion, relative to its absence,

[11] Recall that, in our model, and for the chosen parametrization in our experiments, the firm always wishes the worker to choose the effort level e_H.

and (ii) easier to satisfy in treatment T1S compared to T1N (this follows from (3.7)). Thus, we would expect effort to be higher in T1S relative to T1N and T0 (Proposition 4 below). We are only interested in situations where the ICC is violated. The condition in (3.8) is violated if

$$\Delta u < \frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p};\ k \in \{T1N, T1S\}. \tag{3.9}$$

Suppose that the worker is risk neutral over small stakes ($\Delta u = u(w_1) - u(w_0) \approx w_1 - w_0$). In the presence of guilt aversion, and using (3.9), the ICC is violated if

$$w_1 < w_0 + \frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p} = a_1 - \frac{\lambda_k \phi(\Delta e)}{\triangle p} \equiv a_2, \tag{3.10}$$

where a_1 is defined in (3.4) for treatment T0. It can be readily shown, along the lines of Proposition 1, that the violation of ICC in (3.10) under risk neutrality ensures that the ICC is violated for all risk averse workers too; we omit the similar proof.

Proposition 2. *Consider risk neutral workers and the ICC in treatment T0 is violated, i.e., (3.4), holds. Let a_1, a_2 be defined respectively, in (3.4) and (3.10) and let the parameter of guilt aversion $\lambda_k > 0$.*[12] *When $a_2 < w_1 < a_1$, then:*
(i) The ICC is violated in treatment T0 and the worker chooses the low effort level e_L.
(ii) The ICC holds in treatment T1 and the worker chooses the high effort level, e_H.
(iii) The range of values of w_1 for which the condition $a_2 < w_1 < a_1$ holds and the worker exerts high effort in treatment T1, is strictly increasing in λ_k; non-decreasing in Δe_L (strictly increasing if $\phi' > 0$); and strictly decreasing in $\triangle p$.

Discussion of Proposition 2: From Proposition 2(i),(ii), even when the classical principal-agent model predicts that the worker will choose a low effort level (T0), guilt-aversion might induce the choice of a higher effort level (T1). This can be empirically tested. For any other parameter values (other than $a_2 < w_1 < a_1$) it is never the case that effort in T0 is higher than T1. Proposition 2(iii) identifies the causal factors that make it more likely that the high effort level is chosen in treatment T1 despite the violation of the ICC in treatment T0 in the classical analysis. A more guilt-averse worker (higher λ_k) is more likely to choose a higher effort level. A greater difference in the two effort levels (higher Δe_L) induces even higher guilt aversion, making it more likely that the worker will choose a higher effort level. We do not have individual specific estimates of the guilt aversion parameter, λ_k, but a choice of e_H in Treatments T1N, T1S, when (3.4) holds, is consistent with the presence of guilt-aversion.[13]

The main implication of Proposition 2 is that we should expect a larger percentage of our subjects to choose the high effort level in Treatments T1S, T1N as compared to Treatment T0. Furthermore, if $\phi' > 0$, then an increase in the size of Δe_L should produce a strict increase in the percentage of subjects who choose the higher effort level in Treatment T1. Otherwise, if $\phi' = 0$, then an increase in Δe_L should not produce higher effort. By contrast, there should be no effect on optimal effort of a variation in Δe_L in the classical principal-agent model because if the ICC fails, as in (3.4), the worker should always pick the lower effort level.

[12] If $\lambda_k = 0$, i.e., guilt-aversion is absent, then the results are as in treatment T0.
[13] As noted in Remark 1, workers might also choose high effort on account of guilt-aversion that arises from falling below internal moral norms of behavior. This applies to all treatments in our model.

Our qualitative results extend to the case of risk aversion, as shown in the next proposition.

Proposition 3. The qualitative results in Proposition 2 hold in the presence of risk aversion.

In the next proposition, we examine the treatment contrasts.

Proposition 4. *Suppose that the parameter of guilt aversion is strictly positive,* $\lambda_k > 0$.
(i) High effort is more likely in treatments T1N and T1S relative to treatment T0.
(ii) High effort is more likely in treatment T1S relative to treatment T1N.

Discussion of Proposition 4: Workers who suffer from guilt-aversion exert higher effort to meet expectations arising from the firm culture of high effort (Proposition 4(i)). Furthermore, sanctions by the firm for not meeting the effort standards expected under firm culture, heighten guilt-aversion and induce higher effort since $\lambda_{T1N} < \lambda_{T1S}$ (Proposition 4(ii)).

3.3 Treatment 2: Shame aversion in principal-agent contracts

Treatment 2 incorporates the role of industry-wide social norms of high effort, underpinned by shame-aversion that are external to the firm. We vary the "strength" of the social norms. Unlike guilt, which arises from falling below the expectations of the firm, shame arises from falling below the expectations of one's social or peer group.

Successful social norms require the satisfaction of three key conditions that we have outlined in Section 1.3 in the introduction (normative expectations; empirical expectations; consistency between normative and empirical expectations; and social sanctions). In our experimental design, we ensure that these conditions are met. If these conditions are met, then shame-averse workers are likely to experience shame from falling below the effort expectations of their social/peer group.

The sequence of moves in Treatment 2 is identical to T0 except that there is a Stage 0 that precedes Stage 1 and a Stage 4 that follows Stage 3.

1. Stage 0: Workers learn that $x\%$ of the members of the social group, who have played a similar game before, stated that the worker "ought" to choose the effort level e_H.[14] The worker is also told that the percentage of members of the social group who choose the effort level e_H, when they played the game, is close to $x\%$.[15]

2. Stage 4: The social group can express disapproval of low effort choices of the workers.

The worker suffers a shame-aversion cost that is non-decreasing in the shortfall in effort relative to the expectations of the social/peer group. Consider the shame-aversion function $s : \mathbb{R} \to \mathbb{R}$:

$$s(e_j) = \psi(\Delta e_j),\ j = L, H, \tag{3.11}$$

[14]The term "ought to" is a form of normative injunction that is critical to the formation of normative expectations. We are only interested in 'situations' where the social norm is to exert a high effort level, but we vary the strength of the social norm by varying the support there exists among the population for the high level of effort, i.e., the level of $x\%$ in different treatments (T2H, T2L). We neither elicit normative support for the low effort in our experiments, nor do we provide any information to the subjects about the normative support for low effort.

[15]This ensures that the empirical and normative expectations are aligned.

where $\Delta e_j = e_H - e_j$; $j = L, H$. ψ is a non-decreasing function, $\psi' \geq 0$, such that $\psi(0) = 0$, thus, the worker suffers no shame when choosing an effort level e_H. In the presence of shame-aversion, the worker's utility function from choosing the effort level e_j, is given by

$$W(e_j) \equiv W_j = [(1-p_j)u(w_0) + p_j u(w_1) - c_j] - \mu(x)s(e_j)\,;\; j = H, L, \tag{3.12}$$

where $\mu(x) \geq 0$ the relative weight given to shame-aversion, is increasing in the percentage, x, of other members of the social group who expect group members to exert the high effort level, e_H, so $\mu' > 0$. It follows that the stronger is the normative injunction (i.e., the higher is x), the higher the shame-aversion that is felt by the worker from violating the norm. Comparing the objective functions in treatments T0 and T2, (2.2) and (3.12), the only difference is the presence of the last term in (3.12) that captures the disutility from shame aversion by exerting an effort level lower than the normative expectation of the social group. The two utility functions are identical for a worker who suffers no shame-aversion, i.e., $\mu = 0$. Since $\psi(0) = 0$, the utility from the effort level e_H continues to be given as in treatment T0 by (2.3), so $W_H \equiv U_H$.

The incentive compatibility condition in the presence of social norms is

$$ICC : W_H > W_L \Leftrightarrow (\triangle p)\,\Delta u > \Delta c - \mu(x)\psi(\Delta e_L). \tag{3.13}$$

The ICC in (3.13) is violated if

$$\Delta u < \frac{\Delta c - \mu(x)\psi(\Delta e_L)}{\triangle p}. \tag{3.14}$$

Suppose that the worker is risk neutral over small stakes ($\Delta u = u(w_1) - u(w_0) \approx w_1 - w_0$). In the presence of shame-aversion, and using (3.14), the ICC is violated if

$$w_1 < w_0 + \frac{\Delta c - \mu(x)\psi(\Delta e_L)}{\triangle p} = a_1 - \frac{\mu(x)\psi(\Delta e_L)}{\triangle p} \equiv a_3, \tag{3.15}$$

where a_1 is defined in treatment T0 in (3.4). It can be readily shown, along the lines of Proposition 1, that the satisfaction of the inequality in (3.15) will ensure that the ICC is also violated for all risk-averse workers too; we omit the simple proof.

Proposition 5. *Suppose that the worker is risk neutral and the ICC in treatment T0 is violated, i.e., (3.4) holds. Let a_1, a_3 be defined respectively, in (3.4) and (3.15) and let $\mu > 0$. When $a_3 < w_1 < a_1$, we get the following results.*
(i) The ICC in treatment T0 is violated and the worker chooses the low effort level e_L.
(ii) The ICC is satisfied in treatment T2 and the worker chooses the high effort level e_H.
(iii) The range of values of w_1 for which the condition $a_3 < w_1 < a_1$ holds, and the worker exerts high effort in T2, is strictly increasing in x; non-decreasing in Δe_L (strictly increasing if $\psi' > 0$); and strictly decreasing in $\triangle p$.

From Proposition 5(i),(ii), even when the classical principal-agent model in treatment T0 predicts the choice of a low effort level, social norms, underpinned by shame-aversion in treatment T2, induce the choice of a higher effort level. If the condition in Proposition 5 does not hold ($a_3 < a_1 < w_1$, $w_1 < a_3 < a_1$) it is easily checked that the effort level in treatment T2 is

never lower than T0. Thus, effort is always higher in T2 relative to T0. This can be empirically tested. Proposition 5(iii) identifies the causal factors that make it more likely that the high effort level is chosen in treatment T2 despite the violation of the ICC in the classical analysis. Greater shame-aversion (higher μ and/or higher Δe_L) makes it more likely that the worker will choose a higher effort level. If the proportion of the social group, x, advocating a high normative injunction in favor of a norm of high effort, is higher, then shame-aversion increases. Hence, higher effort is more likely because the ICC is more likely to be satisfied. We use two different values of x to test this prediction in our experiments: Treatment T2H (high value of x) and treatment T2L (low value of x). Thus, we expect high effort to be more likely under treatment T2H.

The results in Proposition 5 also hold qualitatively for risk averse workers; the proof is analogous to the proof of Proposition 3, hence, it is omitted.

Proposition 6. *(i) The worker is more likely to exert higher effort in treatment T2 relative to treatment T0. (ii) The gap in high effort level between treatments T2 and T0 is likely to be higher when x is higher; in other words this gap is higher in T2H relative to T2L.*

From Proposition 6, the gap in high effort between treatments T2H (high x) and T0 is likely to be higher than the gap between T2L (low x) and T0, which can be directly tested in experiments by varying the value of x.

We cannot, however, predict which of the two cases, internal firm norms underpinned by guilt aversion (treatments T1N, T1S), or external firm norms underpinned by shame aversion (treatments T2H, T2L) will produce a higher effort level. This is an empirical question. From Proposition 2, in the interval $a_2 < w_1 < a_1$, subjects will choose the higher effort level under firm culture relative to the baseline treatment T0. From Proposition 5, in the interval $a_3 < w_1 < a_1$, subjects will choose higher effort level under social norms relative to the baseline treatment T0. Thus, the relative efficacy of the two effects depends on the relative sizes of a_2 and a_3. From (3.10) and (3.15), respectively, we have that

$$a_2 \gtreqless a_3 \Leftrightarrow \lambda_k \phi\left(\Delta e\right) \lesseqgtr \mu(x)\psi(\Delta e_L), \tag{3.16}$$

where λ_k and $\mu(x)$ are, respectively, the individual-specific guilt-aversion and shame-aversion parameters that are likely to be heterogeneous across subjects in our experiments. From Proposition 5(iii), a greater proportion of the social group that gives the normative injunction to exert the high effort level (high x) is more likely to ensure that $a_3 < a_2$ because $\mu' > 0$. Thus, social norms of higher effort, as in treatment T2H, are more likely to produce higher effort relative to a reliance on guilt aversion alone (treatments T1N, T1S).

3.4 Optimal effort under fixed wages

Central to classical principal-agent theory is the tradeoff between insurance and incentives. The main insight is that a fixed wage provides insurance but no incentives to choose e_H over e_L. A variable wage, on the other hand, provides poorer insurance, but stronger incentives.

3.4.1 An analysis of optimal effort under fixed wages

Recall that we allow workers to be risk averse in our experiments. Since we choose the contractual parameters so that for risk neutral workers, the ICC must fail, it also fails for risk averse workers (Proposition 1). In the analysis below, we continue with the case of risk neutrality over small stakes to derive the behavioral parameter values for which the ICC fails. For the case of fixed wages, Proposition 7 gives results similar to Propositions 2 and 5 but in terms of cutoff values of the behavioral parameters.

In our fixed wage case, the wage is constant irrespective of the state of the world, good or bad, so that $w_0 = w_1 = \overline{w}$. In particular, in our experiments, we choose the fixed wage $\overline{w}$:

$$\overline{w} = (1 - p_H)w_0 + p_H w_1, \tag{3.17}$$

where w_0 and w_1 are the state dependent wages under variable wages. Thus, the expected profits of the firm are identical under variable and fixed wages and the actions of a risk neutral firm are also unchanged. We are now interested in replicating treatments T0, T1, T2 under fixed wages.

Proposition 7. *Denote the utility differences under fixed wages by* $\Delta u_F = u(\overline{w}) - u(\overline{w}) = 0$ *and under variable wages by* $\Delta u_V = u(w_H) - u(w_L) > 0$. *Suppose that the worker is risk neutral and the ICC in treatment T0 is violated, i.e., (3.4) holds.*
(i) In the classical principal-agent problems in T0, the worker never chooses the high effort.
(ii) Consider treatments T1N, T1S. The ICC always holds under variable wages if it holds under fixed wages. The worker never chooses lower effort under variable wages as compared to fixed wages in treatments T1N and T1S. In particular, if the parameter of guilt-aversion $\lambda_k \in \left(\frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p}, \frac{\Delta c}{\phi(\Delta e)}\right)$ *the worker chooses high effort under variable wages and low effort under fixed wages.*
(iii) Consider treatments T2H, T2L. The ICC always holds under variable wages if it holds under fixed wages. The worker never chooses lower effort under variable wages as compared to fixed wages in treatments T1N and T1S. In particular, if the parameter of shame-aversion $\mu(x) \in \left(\frac{\Delta c - \Delta u_V \triangle p}{\psi(\Delta e_L)}, \frac{\Delta c}{\psi(\Delta e_L)}\right)$ *the worker chooses high effort under variable wages and low effort under fixed wages.*

Discussion of Proposition 7: Under fixed wages, in classical principal-agent models (treatment T0), the worker is fully insured but has no incentives to exert high effort (Proposition 7(i)). In treatments T1N, T1S if the guilt-aversion parameter is high enough, $\lambda_k > \frac{\Delta c}{\phi(\Delta e)}$, the worker chooses high effort for fixed and variable wages. However, for intermediate values of guilt-aversion, $\lambda_k \in \left(\frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p}, \frac{\Delta c}{\phi(\Delta e)}\right)$, the worker chooses high effort under variable wages and low effort under fixed wages. The intuition is that variable wages provide extra incentives to the worker to work harder, loosening the ICC, and ensuring satisfaction of ICC for even lower values of the guilt-aversion parameter. It is never the case that the ICC holds under fixed wages but does not hold under variable wages, hence, effort is always higher under variable wages. An identical intuition holds for why effort is higher under variable wages relative to fixed wages in treatments T2H, T2L (Proposition 7(iii)). In effect, the insurance vs incentives tradeoff in

the classical principal-agent framework serves us well here but it needs to be modified to take account of firm culture and social norms.

3.5 A note on internal moral norms

As noted above in Remark 1, internal moral norms that capture intrinsic motivation might play an important role in effort choice in all treatments. This potentially explains why some workers might choose the high effort in treatment T0 despite the violation of the ICC. However, there may be a conflict between intrinsic and extrinsic motivation (Gneezy and Rustichini, 2000; Bénabou and Tirole, 2003). Several factors might crowd-out, fully or partially, intrinsic motivation, but modelling them formally requires additional machinery and complicates the model. The following two are particularly relevant for us.

1. As noted in the introduction, variable wages in the presence of fixed wage contracts may be interpreted as hostile intent by workers, influencing their behavioral response to guilt and shame (Fehr and Falk, 2002; Dhami, 2019. Vol. 2). Hence, intrinsic motivation from internal moral norms may be crowded-out under variable wages relative to fixed wages.

2. If the worker feels that their autonomy is reduced by the actions of the firms, control-averse workers might interpret it as an unkind action by the firm and withold effort (Deci and Ryan, 1985; Ryan and Deci, 2000; Falk and Kosfeld, 2006). For instance, Proposition 2(iii) shows, that an increase in Δe_L enhances guilt-aversion, and hence the probability of higher effort. However, in treatment T1S the worker faces sanctions, but in T1N there are no sanctions. This might reduce the guilt-aversion of a control-averse worker who might then react inadequately to changes in Δe_L.

In our theoretical model, we can take account of internal moral norms by specifying a function that captures the guilt-cost of choosing low effort, $\omega(e_H - e_L)$, $\omega' \geq 0$, when the internal moral norms dictate the choice of high effort (Freud, 1930/1961; Lazarus, 1991). Thus, in the presence of internal moral norms, the utility function in treatment T0, in (2.2), may be written as

$$EU(e_j) = U_j = (1 - p_j)u(w_0) + p_j u(w_1) - c_j - \kappa\omega(\triangle e); j = L, H,$$

where $\kappa \geq 0$ is the relative importance of internal moral norms. Thus, in the baseline model in T0 augmented with internal moral norms, the ICC in (3.1) would be written as $ICC : U_H > U_L \Leftrightarrow \triangle p\Delta u + \kappa\omega(\Delta e) > \Delta c$; and this is loosened in the presence of internal norms. Thus, if internal moral norms are strong enough in the sense that $\kappa > \frac{\Delta c - \triangle p\Delta u}{\omega(\Delta e)}$, then the worker chooses high effort even if the ICC in the absence of internal norms in (3.1) is violated. We could use the term $\kappa\omega(\Delta e)$ in all treatments in our model to indicate the presence of internal moral norms. However, as argued above, the literature indicates that internal moral norms of this sort may be weakened under variable wages relative to fixed wages. The simplest way to capture this channel is through the following restriction on κ. For $0 \leq \underline{\kappa} < \overline{\kappa}$:

$$\kappa = \begin{cases} \overline{\kappa} & if \quad \text{fixed wage} \\ \underline{\kappa} & if \quad \text{variable wage} \end{cases}.$$

Thus, the effort inducing effects of internal moral norms are higher under fixed wages relative to variable wages.

3.6 A summary of the testable theoretical predictions

Our model predicts the following treatment effects, which essentially constitute our testable hypotheses in the experiments. Note that in every case, the ICC in the classical principal agent model is violated for our parameter values, hence, the classical prediction is that we should observe low effort in all cases.

1. From Proposition 4, we should expect a larger percentage of the subjects to choose the high effort level (i) in treatments T1N and T1S as compared to treatment T0, and (ii) in treatment T1S relative to treatment T1N. From Proposition 6, we should expect a larger percentage of the subjects to choose the high effort level in treatment T2H as compared to treatment T0. From Proposition 5, an increase in the percentage, x, of the social group that gives the normative recommendation to choose high effort, should produce an increase in the percentage of subjects who choose the higher effort level in Treatment 2. Thus, effort is predicted to be higher in treatment T2H relative to treatment T2L.

2. Economic theory cannot a-priori predict which of the two factors (i) firm culture supported by guilt-aversion or (ii) external industry wide norms of effort supported by shame-aversion, plays a stronger role in enhancing effort; see Section 3.3. Hence, one cannot a-priori predict if effort in treatments T2H, T2L will be higher/lower as compared to treatments T1N and T1S. However, from Proposition 5, which predicts that effort is higher in T2H relative to T2L, it is more likely that effort is higher in T2H relative to T1N (and possibly T1S, if x is high enough).

3. From Proposition 7, the effort level under variable wages is always predicted to be higher than that under fixed wages.

4. An increase in the size of Δe_L should not decrease the percentage of subjects who choose the high effort level, e_H, in treatments T1N and T1S (Proposition 2) and in treatments T2L and T2H (Proposition 5). But in the first case the transmission channel is through the effect of internal firm culture mediated by guilt-aversion, while in the second case it is through external industry-specific norms mediated by shame-aversion. The relative quantitative effects are likely to be different, and their relative size is an empirical question. Changes in $\Delta e_L = e_H - e_L$ are predicted to produce no change in effort in the classical principal-agent model. However, from Section 3.5, we know that the relative effects of Δe_L on guilt aversion and shame aversion may be muted in cases where extrinsic motivation crowds out intrinsic motivation (e.g., as in variable wages vs fixed wages or where firm culture is accompanied by sanctions, as in T1S relative to T1N).

4 Experimental design

Our lab experiments were conducted in China with 415 students from Nankai University. No subject participated in the experiment more than once. The identity of subjects stayed anonymous and subjects were assured of anonymity of their responses. There were 17 experimental sessions, and there were 20-30 subjects in each session. The average time taken to complete the experiment was around 40 minutes, and the subjects earned, on average, 48 Chinese Yuan (roughly 6.6 US dollars) including the participation fee. All subjects were paid in private after the experiment. The study was pre-registered; see https://doi.org/10.1257/rct.12110-1.0. All material payoffs in the experiment are expressed in tokens that are converted into Chinese Yuan at the end of the experiment at an exchange rate of 1 token = 0.15 Yuan. Additionally, subjects receive 20 Yuan as a show-up fee for participating in the experiment.

The experimental treatments are in a between-subjects design. In the baseline treatment (T0), subjects are randomly assigned to either of two roles, firms or workers. One worker is randomly matched to one firm.[16] Following the exact sequence of moves as in our theoretical model, firms first make one of the following two choices. (i) Offer the contract, which gives them a positive profit, or (ii) exit the experiment with only the participation fee. If the firm chooses to exit the experiment, then the workers do not need to make any choices. On the other hand, if the firm offered the contract, then the workers make one of the following two choices. (i) Choose the contract that is offered, in which case they also need to choose the effort level, which is either 'high' or 'low', or (ii) choose to exit the experiment with their participation fee. Additionally, to obtain the variation in $\triangle e_L = e_H - e_L$ to test Propositions 2(iii) and 5(iii), we fix the high effort level at $e_H = 8$ but use the strategy method to vary the values of low effort levels, $e_L = 3, 5, 7$, so $\triangle e_L \in \{5, 3, 1\}$. Thus, the workers make effort choices for each of the three 'low' effort levels, 3, 5, 7. The high effort level of 8 costs $c_H = 200$ tokens, while each of the low effort levels (3, 5, or 7) costs $c_L = 100$ tokens each because we wished to separately isolate the effects of guilt-aversion and shame-aversion that depend, partly, on variation in $\triangle e_L$, independent of the cost of effort. A high (low) effort level induces the probability of the good state to be 70% (30%), so $p_H = 0.7$ and $p_L = 0.3$.

The worker's chosen effort level is only privately observed by the worker and not observed by the firm (except in treatment T1S). The state of the world, good or bad, is publicly observed by workers and firms. But workers never know the state-dependent profits of the firm (except that profits are higher in the good state as compared to the bad state), which ensures that there are no considerations of other-regarding preferences.

Workers make their effort choices separately under variable wages and fixed wages, which were run in a counterbalanced order. Under *variable wages*, workers are paid state-dependent wages: $w_1 = 400$ tokens in the good state and $w_0 = 200$ tokens in the bad state. Under *fixed wages*, and using (3.17), workers are paid a fixed wage, $\overline{w} = 340$ tokens, which has the same

[16] Since we are only interested in the worker's effort choices, four workers are randomly matched with one firm and the four workers are totally independent, interact independently with the firm, and have no mutual economic dependence. Dhami et al. (2023) compared the matching of one firm with four independent workers and that of one firm with only one worker, and the results are similar. There is no subject deception.

expected value as the wages under variable wages when $p_H = 0.7$; recall that the firm wishes to implement only the high effort level.

From (3.4), the ICC is violated in the classical principal-agent model for a risk neutral worker if $w_1 < w_0 + \frac{\Delta c}{\triangle p}$. For our parameter values, we have $w_1 = 400$, $w_0 = 200$, $\triangle p = 0.4$, $\Delta c = 100$ so $w_0 + \frac{\Delta c}{\triangle p} = 200 + \frac{100}{0.4} = 450 > w_1 = 400$. This, implies that the ICC is also violated for a risk-averse worker (Proposition 1).

The workers make effort choices only once for the fixed and variable wage case. They are informed of the outcomes (e.g., realization of the good or bad state of the world) only on completing their choices in "both" cases (variable and fixed wages). The worker's income in tokens is calculated separately in the fixed and variable wage cases after the random outcome of their effort is known; this also determines the profits of the firm in tokens. After the experiment, only one case was randomly chosen to pay the subjects. This completes the essential elements of treatment T0. These elements are repeated in the other treatments (T1N, T1S, T2H, T2L). We now describe the additional features of these treatments.

Treatment 1 has two sub-treatments, T1N and T1S, which unlike the baseline treatment T0, make salient the firm culture of high effort, e_H, to the workers. The workers in T1N were informed that "We, the firm subscribes to and takes pride in some core values. One of these core values is to choose a high effort level equal to 8, as compared to a low effort level that is less than 8. We expect workers to put in a high effort level, although we do not monitor the effort of workers, and as such cannot impose any penalties for low effort." The difference between the treatments T1S and T1N is that in T1S the worker's effort choice is observed by the matched firm and the firm can express disapproval of the choice of a low effort by showing a red thumbs down sign on the computer screen to the worker for 3 minutes (non-monetary punishment).[17] This form of punishment makes treatment T1S comparable to treatments T2L, T2H in which the relevant social group can disapprove low effort choices of the workers in the presence of social norms. This allows us to directly compare the relative effects of firm culture mediated by guilt-aversion (treatment T1S) and social norms mediated by shame aversion (treatments T2H, T2L).

In treatment T2, the workers are given two signals of normative expectation and the strategy method is used to make effort choices for each signal. Recall that in treatment T2, x is the percentage of the social group that gives the normative injunction to undertake higher effort level; we have $x = 52\%$ (subtreatment T2H) and $x = 30\%$ (subtreatment T2L). The information on normative injunctions was conveyed to the workers as follows "$x\%$ of your social group believes that workers OUGHT to put in a high effort of 8." Workers then make their effort decisions under the high and low normative signals in the respective treatments T2H and T2L. Workers also receive the following signal of empirical expectation: "50% of your social group chose the high effort level of 8 in similar experiments previously".[18] If a worker falls short of

[17] In contrast to the instructions under T1N, the instructions under T1S specified the following: "The firm can observe the effort level of the worker. But the firm cannot impose any monetary punishments on the worker for low effort. Nor can the firm take any legal action against the worker based on this information." But workers were informed that the firm in treatment T1S shall express disapproval of low effort level through a red thumbs down sign for 3 minutes.

[18] The value of the signal of the empirical expectation (50%) is from the workers' actual effort choices in T0

the high effort expectations of their social group, they are sanctioned by the social group with a probability of 8%.[19] Sanctions are non-monetary and take the same form as in treatment T1S to ensure comparability across treatments (red thumbs down sign on the computer screen to the worker for 3 minutes).

Table 1: Experimental treatments.

Treatment		Information	No. of subjects
T0		no signal	100
T1	T1N	signal of firm culture	100
	T1S	signal of firm culture & firm sanctions	100
T2	T2H	high signal of social norm & social sanctions	115
	T2L	low signal of social norm & social sanctions	115

Table 1 gives information on the number of workers in each treatment. There are 100 workers each in treatments T0, T1 and 115 in treatment T2. The same subjects in treatment T2 play the two subtreatments T2L and T2H in a strategy design because they are identical in all respects except for the signal of normative expectations, $x = 52\%, 30\%$. However, within treatment T1, different subjects play the two subtreatments T1S and T1N.

The independent variables are as follows (other variables are defined as needed).

Variable wage: Dummy variable that equals 1 for variable wages, and 0 for fixed wages.

Age: Subject's age.

Male: Dummy variable that equals 1 for male subjects and 0 otherwise.

Business: Dummy variable that equals 1 for business/economics subjects, and 0 otherwise.

Experience: Dummy variable that equals 1 if the subjects have attended similar experiments before, and 0 otherwise.

Income: Subject's annual household income.

5 Experimental results

In Section 5.1, we provide the basic descriptive statistics on treatment contrasts without conditioning on the control variables. The analysis with controls is conducted in Section 5.2.

5.1 Unconditional descriptive statistics

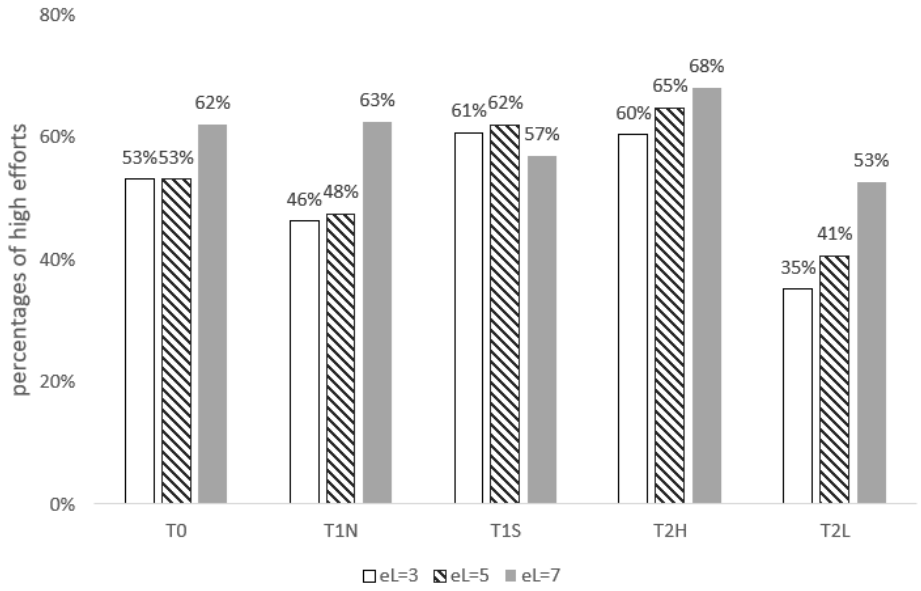
5.1.1 Binary treatment contrasts

Figure 1 reports the unconditional treatment differences in the percentage of subjects who chose the high effort level in each treatment, separated by the variable wage case and the fixed wage

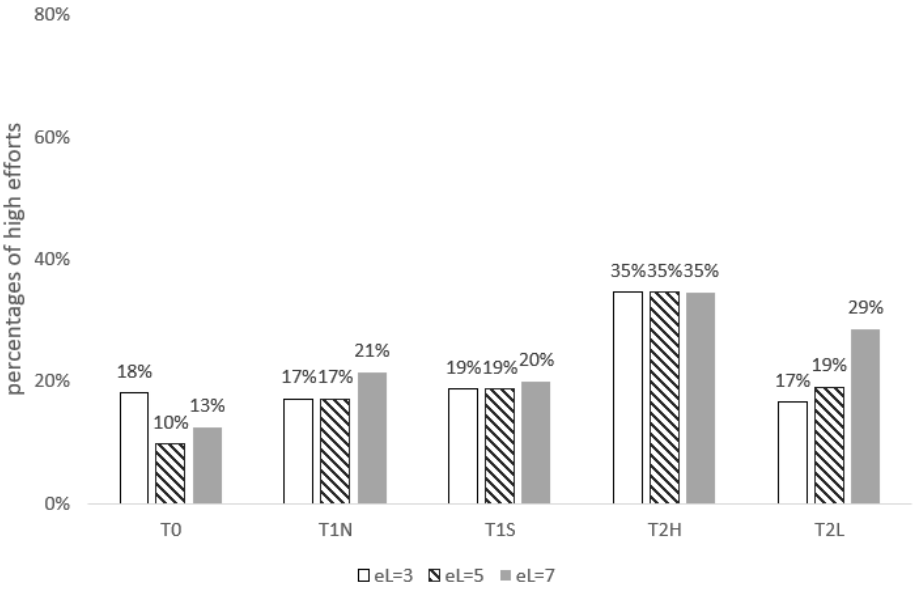
and T1. The high and low signals of normative expectation, 52% and 30%, are from the answers in the post-experimental survey questions in T0 and T1: "Do you think the workers in similar experiments OUGHT TO put in a high effort? (Yes/No)". The source of the normative signals is irrelevant for our theory, so long as the signals provide credible information from the social group. This particular method was used to avoid the charge of subject deception that might have arisen from hypothetical experimenter-constructed normative signals, although those would have allowed us to study a wider range of normative expectations.

[19] The sanction probability (8%) is from the answers of the post-experimental survey question in T0 and T1: "If you could observe a worker's effort to be low in similar experiments, would you disapprove of it? (Yes/No)."

case; see the panel (a) and (b). Table 2 provides a summary of the unconditional pairwise treatment differences in Figure 1 in terms of the presence/absence of statistical significance. Where the differences are not statistically significant, Table 2 reports "no significant difference." However, when we introduce controls and interaction effects in probit regressions in Section 5.2, we find a more nuanced effect of treatment differences in all contrasts considered in Table 2.



(a) Variable wage case



(b) Fixed wage case

Figure 1: Percentages of High Effort Choices.

In both the variable and fixed wage cases, the percentages of high effort choices in T0 are not significantly different from those in T1N and T1S at all the three low effort levels $e_L = 3, 5, 7$ (two-sided z test, $p > 0.1$ in all cases). However, the percentage of high effort choices in T1S

Table 2: Percentage of high effort choices: Pairwise treatment differences.

Treatment contrast	Variable wage	Fixed wage
T0 vs T1N & T1S	no significant difference	no significant difference
T0 vs T2H	no significant difference	T2H>T0 for $e_L = 3, 5, 7$
T0 vs T2L	no significant difference	no significant difference
T1S vs T1N	T1S>T1N for $e_L = 5, 7$	no significant difference
T2H vs T2L	T2H>T2L for $e_L = 3, 5, 7$	T2H>T2L for $e_L = 3, 5, 7$

are greater than those in T0 at all the three low effort levels $e_L = 3, 5, 7$ (except for $e_L = 7$ in the variable wage case): $61\% > 53\%$, $62\% > 53\%$ in the variable wage case, and $19\% > 18\%$, $19\% > 10\%$, $20\% > 10\%$ in the fixed wage case. Thus, the differences between T0 and T1S have the predicted sign, but lack statistical significance.

Our model predicts that high effort is more likely in T1S (firm culture with sanctions) than T1N (firm culture without sanctions). Figure 1 shows that, in the variable wage case, the high effort percentages in T1S are greater than those in T1N, and significant at $e_L = 5, 7$ (two-sided z test, $p < 0.1$). However, in the fixed wage case, these differences are not significant (two-sided z test, $p > 0.1$). We examine the potential reasons in our conditional analysis in Section 5.2.

Next, we test the prediction that a high effort is more likely under treatment T2H as compared to the baseline treatment, T0 (Proposition 6). Our high signal of normative expectations, $x = 52\%$, in treatment T2H, taken from an actual experiment, is relatively moderate, hence, we probably understate the effects of industry-wide social norms. From Figure 1, in the variable wage case, the percentage of high effort choices in T2H at all low effort levels ($e_L = 3, 5, 7$) is greater than that in T0 but the difference is statistically insignificant (two-sided z test, $p > 0.1$). On the other hand, in the fixed wage case, the percentage of high effort choices in T2H at all low effort levels ($e_L = 3, 5, 7$) is significantly greater than that in T0 (one-sided z test, $p < 0.05$); and the percentage of high effort choices under weak social norms in treatment T2L is either more or less than that in T0 but insignificant (two-sided z test, $p > 0.05$). Thus, relative to the baseline treatment, social norms enhance effort when the social norms are strong enough (T2H), but not when social norms are weak (T2L). Furthermore, the contrast between T2H and T0 attains statistical significance under fixed wages. Effort is not lower under T2H in any of the cases.

Comparing the percentage of high effort choices in treatments T2H and T2L for different values of e_L in the variable wage case, we get $60\% > 35\%$ at $e_L = 3$, $65\% > 41\%$ at $e_L = 5$, and $68\% > 53\%$ at $e_L = 7$. The corresponding comparisons for the fixed wage case are $35\% > 17\%$ at $e_L = 3$, $35\% > 19\%$ at $e_L = 5$, and $35\% > 29\%$ at $e_L = 7$. In other words, the percentage of high effort choices in T2H is significantly higher than that in T2L at each of the three low effort levels in both the variable and fixed wage cases (one-sided z test, $p < 0.05$)[20]. This confirms our prediction (Proposition 5(iii), Proposition 6). Thus, social norms underpinned by shame-aversion play an important role in worker's decisions.

Economic theory cannot predict which of the two emotions, guilt and shame, plays a stronger

[20]The only insignificant case is the fixed wage case at low effort 7.

role in enhancing worker's effort. To answer this question, we examine which treatment (T1 or T2) has a greater percentage of high effort choices. Let us first compare T1N (firm culture with no sanctions) with the two treatments T2L and T2H:

(i) In both the fixed and variable wage cases the percentage of high effort choices is significantly greater in T2H relative to T1N at each value of $e_L = 3, 5, 7$.[21]

(ii) The percentage of high effort choices is not significantly different between T2L and T1N at each value of $e_L = 3, 5, 7$ in both the fixed and variable wage cases (two-sided z test, p-values> 0.1). This is expected because the normative injunction for high effort in treatment T2L, $x = 30\%$, was "too low"; i.e., the social norm was too weak.

We now compare T1S (firm culture with sanctions) with the two treatments T2L and T2H. In particular, the differences between variable and fixed wage cases are explained with our more detailed conditional analysis in Section 5.2.

(i) The percentage of subjects choosing high effort in the fixed wage case is significantly greater in T2H relative to T1S at each value of $e_L = 3, 5, 7$ (one-sided z test, $p < 0.05$). There are no significant differences in the variable wage case. Thus, under fixed wages, when social norms are relatively strong, shame-aversion plays a stronger role relative to guilt-aversion.

(ii) In the variable wage case, the percentage of subjects choosing high effort is greater in T1S relative to T2L at each value of $e_L = 3, 5, 7$, and significant at $e_L = 3, 5$ (one-sided z test, $p < 0.01$). Hence, when the social norms are weak, guilt-aversion, particularly when it is heigtened through sanctions by the firm, plays a more important role relative to shame under variable wages. There are no significant differences in the fixed wage case.

5.1.2 Testing the unconditional effects of effort differences, $\triangle e_L$

As noted above, our theoretical model predicts that an increase in the size of $\triangle e_L = e_H - e_L$ should not decrease the percentage of subjects who choose the high effort level, e_H, in treatments T1N and T1S (Proposition 2) and in treatments T2H, T2L (Proposition 5). Recall that $e_L \in \{3, 5, 7\}$ and $e_H = 8$, thus $\triangle e_L \in \{1, 3, 5\}$.

Table 3 calculates the frequencies and proportions of workers who chose high effort level e_H at *all* three levels of $\triangle e_L$ (the 3rd column titled e_H); chose low effort level e_L at *all* three levels of $\triangle e_L$ (the 4th column titled e_L); switched once from e_H to e_L for one of $\triangle e_L \in \{1, 3, 5\}$ (the 5th column titled $e_H \to e_L$); switched once from e_L to e_H for one of $\triangle e_L \in \{1, 3, 5\}$ (6th column titled $e_L \to e_H$); and all the other cases where the effort switched back and forth with no clear pattern as we varied $\triangle e_L \in \{1, 3, 5\}$ (7th column titled Others). The switching point in the 5th, 6th, and 7th columns could be at $\triangle e_L = 3$ or $\triangle e_L = 5$ and varied from subject to subject.

Since the ICC is designed to fail in all cases in our experiment, the classical principal-agent model predicts that 100% of the choices should be in the 4th column, where workers choose the low effort e_L at all levels of $\triangle e_L$. Strict conformity with the classical model under variables wages ranges from 22% to 37% and under fixed wages it ranges from 52% to 79% depending on

[21] The only insignificant case is the variable wage case at $e_L = 7$, (two-sided z test, p-values> 0.1). But its direction is consistent that the percentage of high effort choices is greater in T2H relative to T1N.

Table 3: Effort choices.

Treatment		e_H	e_L	$e_H \to e_L$	$e_L \to e_H$	Others
T0	Variable	38% (30/79)	27% (21/79)	20% (16/79)	**11% (9/79)**	4% (3/79)
	Fixed	7% (5/72)	79% (57/72)	3% (2/72)	**8% (6/72)**	3% (2/72)
T1N	Variable	36% (29/80)	28% (22/80)	26% (21/80)	**10% (8/80)**	0% (0/80)
	Fixed	10% (7/69)	71% (49/69)	12% (8/69)	**7% (5/69)**	0% (0/69)
T1S	Variable	41% (32/79)	24% (19/79)	15% (12/79)	**19% (15/79)**	1% (1/79)
	Fixed	9% (7/80)	69% (55/80)	11% (9/80)	**10% (8/80)**	1% (1/80)
T2H	Variable	48% (44/91)	22% (20/91)	18% (16/91)	**10% (9/91)**	2% (2/91)
	Fixed	21% (18/84)	52% (44/84)	13% (11/84)	**14% (11/84)**	0% (0/84)
T2L	Variable	26% (24/91)	37% (34/91)	24% (22/91)	**8% (6/91)**	5% (5/91)
	Fixed	7% (6/84)	62% (52/84)	20% (17/84)	**9% (7/84)**	2% (2/84)

the treatment.[22]

The percentage of all e_H choices in column 3 might not appear to be too high (although under variable wages nearly 40% of workers on average choose high effort). However, these choices must be seen in the context that the ICC fails. Yet, the treatment contrasts and results based on probit analyses of effort choice give us important information about human behavior in this class of models. For the case of variable wages, and with the exception of treatment T2L where social norms are weak, the percentage of workers choosing e_H, for all values of $\triangle e_L$, rather than e_L is relatively higher (compare the 3rd and 4th columns).

In the baseline treatment, T0, in the absence of firm culture and social norms, 27% of the workers in the variable wage case and 79% in the fixed wage case always choose the low effort; the rest violate the predictions of the classical principal-agent model and 38% of the subjects in the variable wage case always choose the high effort. What accounts for the choice of high effort in treatment T0? There are two potential explanations.

(1) As argued in Remark 1 and Section 3.5, a subset of the workers might have internal moral norms of high effort that do not require an external disciplining device (Bicchieri, 2006; Elster, 2011). This is also related to the idea of intrinsic motivation (Bénabou and Tirole, 2003) and choosing the action that is perceived to be morally correct (Gintis, 2017; Cappelen et al., 2023).

(2) Conditional reciprocity, as in gift exchange games, that motivates workers to reciprocate

[22] In any column where subjects choose the high effort for at least one value of $\triangle e_L \in \{1, 3, 5\}$ is not strictly consistent with the classical predictions.

the gift of a high wage by exerting higher effort.[23]

In Table 3, an average of 9.6% of the workers in the fixed wage case and 11.2% in the variable wage case switched from low effort to high effort as $\triangle e_L$ increases. The behavior of these workers is consistent with our predictions (Proposition 2(iii), Proposition 5(iii)). However, the behavior of workers who switch from high to low effort as $\triangle e_L$ increases is puzzling, particularly since all levels of low effort, $e_L \in \{3, 5, 7\}$, have an identical cost of effort. We offer two potential explanations. (i) Subject miscalculation and error where subjects mistakenly use real life experiences that associate low effort with lower cost of effort. By contrast, in our experiments all low levels of effort, $e_L \in \{3, 5, 7\}$, have an identical effort cost of 100 tokens, which might have been counter-intuitive to some subjects. (ii) As the gap $\triangle e_L$ increases and the firm asks for a high effort level of e_H, some worker's might be aversive to control and view it as a reduction in autonomy (Deci and Ryan, 1985; Ryan and Deci, 2000; Falk and Kosfeld, 2006). Our data cannot distinguish between these alternative explanations.

Consider a comparison between the variable and fixed wage cases in the treatments T1N, T1S, T2H, and T2L. Relatively more workers in the variable wage case exert high effort and relatively more workers exert low effort in the fixed wage case; this is consistent with Proposition 7. These results are also consistent with the literature that workers exert higher effort under variable wages than fixed wages, although the ICC is not violated in this literature (Eriksson and Villeval, 2008; Lazear, 2000; Paarsch and Shearer, 2000).

Table 3 also shows that more workers exerted high effort at *all* three values of $\triangle e_L = 1, 3, 5$ in treatment T2H as compared to T2L.

5.2 Determinants of the probability of high effort choices

We separate the analysis into the effects of firm culture (treatments T1N, T1S; Table 4) and social norm (treatments T2H, T2L; Table 5).

5.2.1 Comparing treatments T0 and T1 (firm culture)

Model 1 of Table 4 contrasts treatments T0 and T1S, while Model 2 contrasts treatments T0 and T1N. The dummy variable *Treatment* equals 1 if the data is from T1S (Model 1) or T1N (Model 2), and 0 if from T0. Table 4 shows the marginal average effects of the Probit models, i.e., the change in the probability that a worker chooses high effort level, when the corresponding independent variable changes by 1 unit. Model 1 and Model 2 are both clustered on experimental sessions.

We now study several comparative static effects, denoting the marginal effects from probit regressions by $\Delta P(e_H \mid S)$, where S is the set of conditions under which we study the marginal effect. Denote the variable wage case by V and the fixed wage case with F (these two cases correspond to setting "Variable wage" in Table 4 equal to 1 and 0 respectively). We use the same terminology in Section 5.2.2 below.

[23] We have not formally modeled conditional reciprocity using models of psychological game theory, but this motive is well understood and documented (Dhami, 2020, Vo. 4; Battigalli and Dufwenberg, 2022).

Table 4: Probability of high effort choices w/o firm culture.

Dependent variable	Probability of high effort choices	
Probit	Model 1	Model 2
Treatment	0.05*	0.18*
	[0.028]	[0.106]
Variable wage	0.53***	0.52***
	[0.063]	[0.060]
$\triangle e_L$	0.02**	0.02**
	[0.010]	[0.010]
Treatment×Variable wage	-0.07***	-0.18
	[0.024]	[0.115]
Variable wage×$\triangle e_L$	-0.04***	-0.04***
	[0.005]	[0.005]
Treatment×$\triangle e_L$	0.02***	-0.04***
	[0.005]	[0.014]
Treatment×Variable wage×$\triangle e_L$	-0.01**	0.03***
	[0.004]	[0.008]
Age	0.02**	0.01
	[0.007]	[0.008]
Male	0.01	-0.01
	[0.071]	[0.048]
Income	-0.02*	-0.00
	[0.013]	[0.020]
Business	0.03	-0.01
	[0.026]	[0.044]
Experience	-0.06	0.03
	[0.052]	[0.041]
Data	T0 vs T1S	T0 vs T1N
No. of Obs.	924	891

Notes: The standard errors (clustered on sessions) are in the brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

1. Keeping the treatments fixed at T1S and T1N, we wish to examine the relative effect of variable wages (V) and fixed wages (F) in each treatment (Model 1):

$$\Delta P(e_H \mid T1S, V) - \Delta P(e_H \mid T1S, F) = 0.53 - 0.07 - 0.04 - 0.01 = 0.41.$$

Thus, in treatment T1S the probability of high effort is 41% higher under variable wages relative to fixed wages. The corresponding difference in (i) treatment T1N is 33% higher ($\Delta P(e_H \mid T1N, V) - \Delta P(e_H \mid T1N, F) = 0.33$), and (ii) treatment T0 is 49% higher in Model 1 ($\Delta P(e_H \mid T0, V) - \Delta P(e_H \mid T0, F) = 0.53 - 0.04 = 0.49$).
Thus, variable wages enhance effort relative to fixed wages, as predicted (Proposition 7). Within each treatment, the classical insurance versus incentives tradeoff plays an important role in explaining why effort is higher under variable wages rather than fixed wages. Variable wages provide better incentives for high effort. But the classical framework cannot explain the effects in (2) and (3) below.

2. Keeping fixed the type of wage, variable wage (V) or fixed wage (F), consider the treatment differences T1S–T0 and T1N–T0 for the marginal probability of high effort.

$$\Delta P(e_H \mid T1S, V) - \Delta P(e_H \mid T0, V) = 0.05 - 0.07 + 0.02 - 0.01 = -0.01. \tag{5.1}$$

$$\Delta P(e_H \mid T1S, F) - \Delta P(e_H \mid T0, F) = 0.05 + 0.02 = 0.07. \tag{5.2}$$

From (5.1), there is virtually no treatment difference between T1S (firm culture with sanctions) and T0 in the variable wage case. However, from (5.2), in the fixed wage case, higher effort is 7% more likely under treatment T1S relative to T0. By contrast, under variable wages, simple calculations show that the treatment difference between T1N (firm culture with no sanctions) and T0 is identical but under fixed wages, it increases to 14%.[24]
Why do we observe higher treatment differences between T1S and T0 (and T1N and T0) for the fixed wage case but not the variable wage case. The crowding-out effect of intrinsic motivation (Section 3.5) is likely to apply to two different dimensions in our model.
(a) We expect guilt-aversion that enhances high effort in treatment T1 (firm culture), to apply to a greater extent under fixed wages relative to variable wages. Extrinsic motivation (in the form of variable wages, relative to fixed wages) can crowd out intrinsic motivation, of which guilt-aversion is an important component (Gneezy and Rustichini, 2000; Bénabou and Tirole, 2003). This potentially explains the treatment differences between T1S and T0 in (5.1) and (5.2) ($7\% > -1\%$) and also the treatment differences between T1N and T0 ($14\% > -1\%$).
(b) In treatment T1S, crowding-out of guilt aversion is likely to be stronger in the presence of sanctions by the firms on workers (which is potentially interpreted as hostile intent on the part of the firm by the workers) relative to treatment T1N where no sanctions are imposed. Recall from (3.4) that the absolute level of guilt aversion is assumed to be higher under T1S relative to T1N, $\lambda_{T1N} < \lambda_{T1S}$, and this explains the treatment differences between T1N and T1S. However, our data is consistent with the following further

[24] We have: $\Delta P(e_H \mid T1N, V) - \Delta P(e_H \mid T0, V) = -0.01$ and $\Delta P(e_H \mid T1N, F) - \Delta P(e_H \mid T0, F) = 0.14$.

interpretation that is implied by our informal discussion on crowding-out of intrinsic motivation in Section 3.5. On account of the crowding out effect, the gap $\lambda_{T1S} - \lambda_{T1N}$ is relatively smaller under variable wages, relative to fixed wages. Hence, under fixed wages, the treatment effects are stronger in T1N vs T0 relative to T1S vs T0 (14% > 7%). We present more supportive evidence in the next point.

3. Consider the effects of $\triangle e_L$ on the probability of high effort. We observe a relatively small direct marginal effect of $\triangle e_L$ on the probability of high effort, about 2% higher effort on average in both models in Table 4 as $\triangle e_L$ increases. This prompts the following interpretation. From (3.6), when workers choose the low effort, e_L, they experience guilt aversion given by $\lambda_k \phi(\Delta e)$, where $\phi' \geq 0$, $\phi(0) = 0$, and k is an index for the two subtreatments T1S and T1N. The small direct effects of $\triangle e$ suggest that the function ϕ is relatively flat ($\phi' \approx 0$), and that subjects largely experience a fixed amount of guilt, λ_k, from violating firm culture. In other words, the data suggests that λ_k is a good approximation to $\lambda_k \phi(\Delta e)$ in the objective function of the worker, (3.6).
As noted in Section 3.5, in the presence of variable wages (as compared to fixed wages) external moral norms of firm culture might crowd-out internal moral norms. We note that the interaction 'Treatment × Variable wage × $\triangle e_L$' is negative in both models and it is statistically significant for Model 1. Thus, in the presence of firm sanctions and variable wages, an increase in $\triangle e_L$ produces a lower effect on the probability of high effort in treatment T1S relative to T0. Thus, it appears that incentives in the form of firm sanctions and variable wages diminish the guilt-response to an increase in $\triangle e_L$. This supports the argument made in 2b in the previous point.

5.2.2 Comparing treatments T0 and T2 (social norms)

In this section, we compare the probability of high effort choices in treatments T0 and T2L, T2H using a Probit model to determine the marginal effect of social norms. Model 1 in Table 5 uses the data of the workers from treatments T0 and T2L, while Model 2 uses the data of the workers from treatments T0 and T2H. The dummy variable *Norm* equals 1 if the data is from T2L (Model 1) or T2H (Model 2), and 0 if the data is from T0. Both models are clustered on experimental sessions.

In Table 5, the marginal effect of the variable 'Norm' is insignificant in Model 1, but significantly positive in Model 2, otherwise the results in the two models are similar.[25] Our discussion below parallels the one in Section 5.2.1, so our explanations will be brief.

1. Keeping the treatments fixed at T2L and T2H, we wish to examine the effect of variable wages (V) versus fixed wages (F):

$$\Delta P(e_H \mid T2H, V) - \Delta P(e_H \mid T2H, F) = 0.51 - 0.17 - 0.03 = 0.31.$$

[25] Recall that our weak social norms in T2L are relatively weak ($x = 30\%$) and our strong social norms in T2H are reasonably moderate ($x = 52\%$). It would have been interesting to consider experimenter-generated hypothetical values of x such as $x = 50\%, 75\%, 90\%$. However, that might have led to charges of subject deception by experimental purists who might have argued that these are not "actual" value.

Table 5: Probability of high effort choices w/o high or low social norm.

Dependent variable	Probability of high effort choices	
Probit	Model 1	Model 2
Norm	0.34	0.24***
	[0.256]	[0.069]
Variable wage	0.46***	0.51***
	[0.062]	[0.064]
$\triangle e_L$	-0.01	0.01
	[0.013]	[0.007]
Norm×Variable wage	-0.21***	-0.17*
	[0.075]	[0.087]
Variable wage×$\triangle e_L$	-0.02*	-0.03***
	[0.009]	[0.010]
Age	0.02	0.01
	[0.009]	[0.011]
Male	-0.06	-0.07
	[0.046]	[0.073]
Income	-0.01	0.01
	[0.013]	[0.018]
Business	0.04	0.05
	[0.034]	[0.054]
Experience	0.01	-0.04
	[0.063]	[0.063]
Data	T0 & T2L	T0 & T2H
No. of Obs.	972	972

Notes: The standard errors (clustered on sessions) are in the brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

$$\Delta P(e_H \mid T2L, V) - \Delta P(e_H \mid T2L, F) = 0.46 - 0.21 - 0.02 = 0.23.$$

Thus, within each treatment, there is a significant increase in the probability of choosing high effort under variable wages as compared to fixed wages, respectively 31% and 23% in treatments T1H and T1L, which is consistent with Proposition 7). Thus, the incentive effects are powerful. However, the marginal effects of the change from fixed to variable wages are stronger under higher social norms, by 8% (31% − 23%). The explanation is along the same lines as discussed in the previous section, Section 5.2.1.

2. Keeping fixed the type of wage, variable wage (V) or fixed wage (F), we now wish to study the treatment differences between treatments T2H and T2L relative to treatment T0. In particular, the treatment differences under T2H and T0 (Model 2) under variable and fixed wages, respectively, are as follows:

$$\Delta P(e_H \mid T2H, V) - \Delta P(e_H \mid T0, V) = 0.24 - 0.17 = 0.07.$$

$$\Delta P(e_H \mid T2H, F) - \Delta P(e_H \mid T0, F) = 0.24.$$

There is a 7% higher probability of high effort in T2H relative to T0 under variable wages, and a 24% higher probability under fixed wages. For treatment T2L, the corresponding figures are 13% and 34%.[26] Thus, as compared to the baseline treatment T0, social norms enhance effort to a relatively greater extent under fixed wages, as compared to variable wage. It would appear to follow that steeper incentives, in the form of variable wages, partially crowd out "intrinsic motivation" relative to fixed wages. This is an extension of the earlier findings in Section 5.2.1 where incentives crowd out "internal motivation." The observed treatment effects between T0 and T2 in the fixed and variable wage cases confirm the classical finding that subjects experience shame from violating social norms, or that they are programmed to conform to normative injunctions (Bicchieri, 2006; Gintis, 2017).

3. In each of the two treatments, the direct marginal effects of $\triangle e$ are relatively small (of the order of 1%). This result is identical to that in Section 5.2.1, and the explanation is similar. Recall that our shame aversion term, given in (3.12), from choosing the low effort is $\mu(x)s(e_L)$ such that $s(e_L) = \psi(\triangle e)$, where $\psi' \geq 0$, $\psi(0) = 0$. The relatively small marginal effects of $\triangle e$ suggest that the function ψ is relatively flat ($\psi' \approx 0$), so subjects experience a fixed amount of shame, $\mu(x)$, from violating the social norm, irrespective of the extent of the transgression (at least for the magnitudes of transgressions considered in our experiments).

   The negative sign and significance of the interaction term Variable wage and $\triangle e$ lends further weight to the crowding-out of intrinsic motivation in the presence of variable wages.

5.3 Order effect and proportions of subjects exiting the experiment

The variable wage and fixed wage cases are run in a counterbalanced order in our experiments, for each subject. The effort choices at the three levels of low effort in each treatment are

[26] $\Delta P(e_H \mid T2L, V) - \Delta P(e_H \mid T0, V) = 0.34 - 0.21 = 0.13$ and $P(e_H \mid T2L, F) - \Delta P(e_H \mid T0, F) = 0.34$

not significantly different across the two orders of games (Mann-Whitney test, p-value> 0.05). The only significantly different effort choices are at $e_L = 7$ in T2L (Mann-Whitney test, p-value$= 0.027$).

Recall that firms and workers in our experiments have an option of offering/accepting the given contract or exiting the experiment with the participation fee. The main findings are as follows. (i) In the variable wage case, there are almost no firms or workers who chose to exit the experiment. (ii) In the fixed wage case, there are more firms and workers (around 10%) who exit the experiment (in T0, T1N, T1S, T2H, and T2L) but there are little treatment differences. More details can be found in the appendix.

6 Conclusions

The incentive compatibility condition (ICC) plays a central role in economics and it is taken as an article of faith. We agree with the importance of the ICC but call for an enrichment of the ICC to take account of psychological and social motivations. Based on extensive evidence, we argue for at least two extensions: (i) Firm culture or internal firm norms, and (ii) industry wide social norms or external firm norms in influencing behavior. We argue that compliance with firm culture is underpinned by guilt-aversion and compliance with social norms is underpinned by shame aversion. There is interest in determining the relative importance of internal and external firm norms. We consider a simple principal-agent relationship with one firm and one worker but moral hazard in effort.

In all cases considered in our model, the ICC fails. Thus, the classical principal-agent model predicts that all workers must choose the low effort level. Compliance with this prediction is better under fixed wages but it is relatively poor under variable wages with less than 30% of the workers conforming. We show that firm culture produces higher effort when firms can also impose non-monetary sanctions, such as disapproval, even if effort is not verifiable to a third party. External social norms are effective when the normative injunction is stronger relative to when it is weaker. Firm culture is more effective than weak external social norms, but strong external social norms are more effective than firm culture. The precise results depend on whether we consider variable or fixed wages. Our data shows that extrinsic motivation in the form of variable wages (relative to fixed wages) crowds out intrinsic motivation arising from internal moral norms of high effort.

7 Appendix

7.1 Proofs of Propositions

Proof of Proposition 1: Suppose that (3.3) holds. By definition $\Delta u = u(w_1) - u(w_0)$. Taking a second order Taylor series approximation of $u(w_1)$ around $u(w_0)$, we get

$$u(w_1) \approx u(w_0) + u'(w_0)\Delta w + \frac{1}{2}u''(w_0)\Delta w^2.$$

$$\Rightarrow \Delta u = u(w_1) - u(w_0) \approx u'(w_0)\Delta w \left[1 - \frac{1}{2}R_A \Delta w\right], \tag{7.1}$$

where $R_A = -\frac{u''(w_0)}{u'(w_0)}$ is the coefficient of absolute risk aversion. From (3.3) and (7.1), the ICC is violated for a risk averse individual if

$$u'(w_0)\Delta w\left[1-\frac{1}{2}R_A\Delta w\right] < \frac{\Delta c}{\triangle p}. \tag{7.2}$$

For a risk neutral individual, $R_A = 0$. Thus, the corresponding condition for the violation of the ICC for a risk neutral individual is

$$u'(w_0)\Delta w < \frac{\Delta c}{\triangle p}. \tag{7.3}$$

Clearly if (7.3) holds, i.e., ICC is violated for a risk neutral worker then (7.2) also holds, i.e., ICC is violated for a risk averse worker; but the converse is false. This method of proof applies to all the treatments in our paper. ■

Proof of Proposition 2: (i) Follows directly from the restriction $w_1 < a_1$ and (3.4).

(ii) Since $a_2 < w_1$, we get from (3.10) that the ICC holds under guilt-aversion.

(iii) From (3.10), we get

$$a_2 < w_1 < a_1 \Leftrightarrow a_1 - \frac{\lambda_k \phi\left(\Delta e\right)}{\triangle p} < w_1 < a_1. \tag{7.4}$$

As λ_k increases, or as $\triangle p$ decreases, the left hand side of the interval becomes strictly smaller. The left hand side of the interval also becomes strictly smaller as Δe_L increases if $\phi' > 0$ and stays unchanged with a change in Δe_L if $\phi' = 0$. Thus, the range of values for which the condition $a_2 < w_1 < a_1$ holds (so the ICC holds in treatment T1 but not in T0) is increasing in λ_k, Δe_L and decreasing in $\triangle p$. ■

Proof of Proposition 3: When the ICC is violated for risk neutral workers in treatment T1 (see (3.10)), the proof of Proposition 1 can be used to show that the ICC is also violated for risk averse workers in treatment T1 and $\Delta u < w_1 - w_0$, so

$$\Delta u < \frac{\Delta c - \lambda_k \phi\left(\Delta e\right)}{\triangle p}. \tag{7.5}$$

Suppose that (7.5) does not hold. We can now choose values of w_0, w_1 and the other parameters of the model such that

$$\frac{\Delta c - \lambda_k \phi\left(\Delta e\right)}{\triangle p} < \Delta u < \frac{\Delta c}{\triangle p}. \tag{7.6}$$

If (7.6) holds, then the worker chooses low effort in treatment T0 and high effort in treatments T1N, T1S. This result is qualitatively identical to Proposition 2(i),(ii). Finally, a qualitatively similar result to Proposition 2(iii) arises because the LHS of (7.6) is decreasing in λ_k and Δe. ■

Proof of Proposition 4: Under the parameter restrictions in Proposition 2, even when the ICC is violated in treatment T0, it is not violated in T1, hence making it more likely that effort is higher in T1 relative to T0. From Proposition 2(iii), the ICC under guilt aversion, $a_2 < w_1$, is more likely to be satisfied when the value of the guilt-aversion parameter λ is higher. From (3.7), $\lambda_{T1N} < \lambda_{T1S}$, hence this implies that the higher effort level is more likely in treatment T1S relative to treatment T1N. ■

Proof of Proposition 5: (i) Follows directly from the restriction $w_1 < a_1$ and (3.4).

(ii) Since $a_3 < w_1$, we get from (3.14), (3.15) that the ICC holds under shame-aversion.

(iii) From (3.15), we get

$$a_3 < w_1 < a_1 \Leftrightarrow a_1 - \frac{\mu(x)\psi(\Delta e_L)}{\triangle p} < w_1 < a_1. \tag{7.7}$$

As x increases, or as $\triangle p$ decreases, the left hand side of the inequality becomes strictly smaller because $\mu' > 0$; and strictly smaller following an increase in Δe_L if $\psi' > 0$. Thus, the range of values for which the condition $a_3 < w_1 < a_1$ holds is increasing in x, Δe_L and decreasing in $\triangle p$. ■

Proof of Proposition 6: (i) Part (i) of the proof is analogous to the proof of Proposition 4, hence, it is omitted (ii) This follows directly from the result in Proposition 5(iii) which shows that in the presence of social norms, high effort is more likely when x is higher. ■

Proof of Proposition 7: Recall that from (3.17), we chose the fixed wage $\overline{w} = (1 - p_H)w_0 + p_H w_1$ that leaves unchanged the decisions of the firm.

(i) From (3.1), and using $\Delta u_F = 0$, the ICC never holds in the classical principal-agent problems in T0 because it requires $0 > \Delta c > 0$. which is impossible. Since workers are fully insured, they choose the lowest effort.

(ii) Consider now treatments T1N, T1S. From (3.8), the ICC holds under variable wages if $\Delta u_V > \frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p}$; $k \in \{T1N, T1S\}$ and under fixed wages if $\Delta u_F = 0 > \frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p}$. Thus, the ICC always holds under variable wages if it holds under fixed wages. But the converse is false. When $0 < \frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p} \Leftrightarrow \lambda_k < \frac{\Delta c}{\phi(\Delta e)}$ the ICC does not hold under fixed wages and the worker chooses low effort. But if $\Delta u_V > \frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p} \Leftrightarrow \lambda_k > \frac{\Delta c - \Delta u_V \triangle p}{\phi(\Delta e)}$ the worker chooses high effort under variable wages. Thus, if $\lambda_k \in \left(\frac{\Delta c - \lambda_k \phi(\Delta e)}{\triangle p}, \frac{\Delta c}{\phi(\Delta e)}\right)$ the worker chooses high effort under variable wages and low effort under fixed wages. Thus, the worker never chooses lower effort under variable wages as compared to fixed wages in treatments T1N and T1S.

(iii) Proceeding as in the proof of (ii), if the ICC holds under fixed wages it always holds under variable wages. If the shame-aversion parameter $\mu(x) \in \left(\frac{\Delta c - \Delta u_V \triangle p}{\psi(\Delta e_L)}, \frac{\Delta c}{\psi(\Delta e_L)}\right)$, then the worker chooses high effort under variable wages but low effort under fixed wages. Otherwise for any other level of the shame-aversion parameter it is never the case that the worker chooses low effort under variable wages if they choose high effort under fixed wages. Hence, the worker never chooses lower effort under variable wages in treatments T2H and T2L. ■

7.2 Proportions of subjects exiting the experiment

In T0, 0%(= 0/20) firms in the variable wage case and 10%(= 2/20) firms in the fixed wage case exited the experiment, and the two proportions are not significantly different (two-sided z test, $p > 0.1$). In T0, 1.25%(= 1/80) workers in the variable wage case and 10%(= 8/80) workers in the fixed wage case exited the experiment, and the two proportions are significantly different (one-sided z test, $p < 0.05$). Similarly, in T1N, 0%(= 0/20) firms in the variable wage case and 10%(= 2/20) firms in the fixed wage case exited the experiment, and the two proportions are not significantly different (two-sided z test, $p > 0.1$). In T1N, 0%(= 0/80) workers in the variable wage case and 12.5%(= 10/80) workers in the fixed wage case exited the

experiment, and the two proportions are significantly different (one-sided z test, $p < 0.05$). In T1S, 0%(= 0/20) firms in the variable wage case and the fixed wage case exited the experiment. In T1S, 1.25%(= 1/80) workers in the variable wage case and 0%(= 0/80) workers in the fixed wage case exited the experiment, and the two proportions are not significantly different (one-sided z test, $p > 0.1$). In T2, 0%(= 0/23) firms in the variable wage case and 8.7%(= 2/23) firms in the fixed wage case exited the experiment, and the two proportions are not significantly different (two-sided z test, $p > 0.1$). In T2, 1.1%(= 1/92) workers in the variable wage case and 8.7%(= 8/92) workers in the fixed wage case exited the experiment, and the two proportions are significantly different (one-sided z test, $p < 0.05$).

References

[1] Akerlof, G. A. (1982). Labor contracts as partial gift exchange. Quarterly Journal of Economics 97(4): 543–69.

[2] Akerlof, G., and Kranton, R. (2000). Economics and Identity. Quarterly Journal of Economics 115(3): 715–53.

[3] Battigalli, P., and Dufwenberg, M. (2022). Belief-Dependent Motivations and Psychological Game Theory. Journal of Economic Literature 60(3): 833-82.

[4] Battigalli, P., and Dufwenberg, M. (2007). Guilt in games. American Economic Review 97(2): 170-176.

[5] Baumeister, R. F., Stillwell, A. M., Heatherton, T. F. (1994). Guilt: An interpersonal approach. Psychological Bulletin 115: 243-267.

[6] Bedford, Olwen (2004). The Individual Experience of Guilt and Shame in Chinese Culture. Culture & Psychology 10(1): 29–52.

[7] Bénabou, R. (2013). Groupthink: Collective Delusions in Organizations and Markets. Review of Economic Studies 80: 429–62

[8] Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. Review of Economic Studies 70(3): 489–20.

[9] Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. American Economic Review 96(5): 1652–78.

[10] Besley, T. and Persson, T. (2022) Organizational dynamics: culture, design, and performance. The Journal of Law, Economics, and Organization 40(2): 1–22.

[11] Besley, T., and Ghatak, M. (2005). Competition and Incentives with Motivated Agents. The American Economic Review 95(3): 616-636.

[12] Bicchieri, C. (2006). The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press: Cambridge.

[13] Bowles, S., and Gintis, H. (2011). A cooperative species: Human reciprocity and its evolution. Princeton University Press: Princeton.

[14] Brown, M., Falk, A., and Fehr, E. (2004). Relational contracts and the nature of market interactions. Econometrica 72(3): 747–80.

[15] Cappelen, A. W., Enke, B., and Tungodden, B. (2023) Universalism: Global Evidence. Forthcoming American Economic Review.

[16] Charness, G. and Dufwenberg, M. (2006). Promises and partnership. Econometrica 74(6): 1579–601.

[17] Chatman, J. A., O'Reilly, C. A., (2016). Paradigm lost: Reinvigorating the study of organizational culture. Research in Organizational Behavior 36: 199–224.

[18] Cyert, R., and March, J. G. (1963). A Behavioral Theory of the Firm. Oxford, UK: Wiley-Blackwell.

[19] Deci, E. L., Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. Journal of research in personality 19(2): 109-134.

[20] Dessein, W., and Prat, A. (2022). Organizational Capital, Corporate Leadership, and Firm Dynamics. Journal of Political Economy 130(6): 1477-1536.

[21] Dhami, S. (2019). The Foundations of Behavioral Economic Analysis. Volume II: Other-Regarding Preferences, Oxford University Press: Oxford.

[22] Dhami, S. (2020). The Foundations of Behavioral Economic Analysis. Volume IV: Behavioral Game Theory, Oxford University Press: Oxford.

[23] Dhami, S., Arshad, J. and al-Nowaihi, A. (2022). Psychological and Social Motivations in Microfinance Contracts: Theory and Evidence. Journal of Development Economics. 158: 102912.

[24] Dhami, S., Mengxing, W., and al-Nowaihi, A. (2023) Classical and Belief-Based Gift Exchange Models: Theory and Evidence. Games and Economic Behavior 138: 171-196

[25] Dhami, S., Wei, M., and al-Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. Journal of Economic Behavior & Organization. 167: 361–390.

[26] Ellingsen, T., Johannesson, M., Tjøtta, S. and Torsvik, G. (2010). Testing Guilt Aversion. Games and Economic Behavior 68(1): 95–107.

[27] Eriksson, T., Villeval, M. C. (2008). Performance-pay, sorting and social motivation. Journal of Economic Behavior & Organization 68(2): 412-421.

[28] Falk, A., Kosfeld, M. (2006). The hidden costs of control. American Economic Review 96(5): 1611-1630.

[29] Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. Quarterly Journal of Economics 108(2): 437–59.

[30] Fehr, E., and Falk, A. (2002) Psychological foundations of incentives. Joseph Schumpeter Lecture. European Economic Review 46: 687–724.

[31] Fehr, E., Klein, A., and Schmidt, K. M. (2007). Fairness and contract design. Econometrica 75(1): 121–54.

[32] Fehr, E., and List, J. (2004) The hidden costs and returns of incentives– trust and trustworthiness among CEOs. Journal of the European Economic Association September 2(5):743–771.

[33] Freud, S. (1961). Civilization and its discontents. New York: Norton. (Originally published 1930).

[34] Gintis, H. (2017) Individuality and entanglement. Princeton University Press.

[35] Graham, J. R., Grennan, J., Harvey, C. R., and Rajgopal, S. (2022) Corporate culture: Evidence from the field. Journal of Financial Economics 146(2): 552-593.

[36] Gneezy, U., and Rustichini, A. (2000) A Fine is a Price. The Journal of Legal Studies 29(1): 1-17.

[37] Greif, A. (1994). Cultural Beliefs and the Organization of Society: A Theoretical and Historical Reflection on Collectivist and Individualist Societies. Journal of Political Economy 102(5): 912–950.

[38] Hofstede, G. (1984). Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations. New York, NY: Sage Publications.

[39] Hofstede, G., Hofstede, G. J., and Minkov, M. (2010). Cultures and Organizations: Software of the Mind. 3rd ed. New York, NY: McGraw Hill.

[40] Holmström, B. (1979). Moral hazard and observability. The Bell journal of economics 10(1): 74-91.

[41] Holmström, B., and P. Milgrom (1991) Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, and Organization 7: 24–52.

[42] Huck, S., Kübler, D., Weibull, J. (2012). Social Norms and Economic Incentives in Firms. Journal of Economic Behavior and Organization 83: 173–185.

[43] Kandel, E., Lazear, E. P. (1992). Peer pressure and partnerships. Journal of Political Economy. 100: 801–817.

[44] Khalmetski, K., Ockenfels, A., Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. Journal of Economic Theory 159: 163-208.

[45] Kitayama, S., Mesquita, B., and Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. Journal of Personality and Social Psychology 91(5): 890–903.

[46] Kreps, D. M. (1990). Corporate Culture and Economic Theory. in James Alt and Kenneth A. Shepsle, eds., Perspectives on Positive Political Economy, Cambridge, UK: Cambridge University Press.

[47] Lazarus, R. S. (1991). Emotion and adaptation. New York: Oxford University Press.

[48] Lazear, E. P. (2000). Performance pay and productivity. American Economic Review 90(5): 1346-1361.

[49] Masclet, D., Noussair, C., Tucker, S., and M.-C. Villeval (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. American Economic Review 93: 366-380.

[50] Mirrlees, James A., (1971). An Exploration in the Theory of Optimal Income Taxation. Review of Economic Studies 38: 175-208.

[51] Paarsch, H. J., Shearer, B. (2000). Piece rates, fixed wages, and incentive effects: Statistical evidence from payroll records. International Economic Review 41(1): 59-92.

[52] Ryan, R. M., Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary educational psychology 25(1): 54-67.

[53] Schein, E. H. (1990). Organizational Culture. American Psychologist 45 (2): 109–19.

[54] Tirole, J. (1996). A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality). Review of Economic Studies 63(1): 1–22..

[55] Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. Econometrica 76(6): 1476–80.

[56] Whyte, William H. 1956. The Organization Man. New York, NY: Simon & Schuster.

[57] Wilson, J. Q. (1989). Bureaucracy: What Government Agencies Do and Why They Do It. New York, NY: Basic Books.