# Pre-Analysis Plan

## The incentive compatibility condition, firm culture, and social norms under moral hazard: Theory and Evidence

**Sanjit Dhami, Mengxing Wei**

# 1. Introduction and Background

We consider a principal-agent model under moral hazard and examine, theoretically, and empirically, the psychological and social motivations of the agent. We examine the effect of such motivations on the incentive compatibility condition (ICC) of the agent. We show that even when the ICC is violated in the classical analysis, if firm culture and social norms are effective, the worker will prefer to exert high effort levels. If this is the case, then the classical analysis overstates the informational rents to workers, as well as the nature of the moral hazard problem.

# 2. Methods

## 2.1. Experimental design overview

There are 3 treatments. It's a between-subjects design.

(1) Baseline treatment (classical principal-agent analysis)

Subjects are randomly assigned to two roles - firms and workers. Each subject is assigned to one and only one of these two roles. There is one firm and one worker in each group.

Subjects in the role of firms 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.

Subjects in the role of 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.

The chosen effort level is only privately observable to the worker, but it is never observed by the firm. The firm cannot observe the effort level of the worker, and hence cannot punish or impose any sanctions on the worker.

Each effort level leads to one of two possible states of the world - a good state and a bad state. (i) A good state of the world, in which the firm receives a high profit, or (ii) a bad state of the world in which the firm receives a low profit.

Information on the level of profits in each state of the world, good state and bad state, is shown only to the firms but not to the workers.

A high (low) effort level by the worker makes the good (bad) state and high (low) profits of the firm, relatively more likely.

Workers make their effort choices for two different cases, both run in a random order.

Variable wage case: In the first case, workers are paid a wage that depends only on the state of the world.

Fixed Wage case: In the second case, workers are paid a fixed wage independent of the state of the world.

The two cases will run only once. After completing responses for one case, subjects cannot know any result from the completed part before starting the second case. After completing both cases, they will get to know the results of both cases. Subject's income in tokens is calculated separately in each case and it depends on their decisions and the decisions of their matched firm or worker. After the experiment, only one case will be randomly chosen to pay the subjects. The identity of subjects stays anonymous.

(2) Treatment 1 (Firm culture, workplace norms, and guilt-aversion)

The only difference than baseline treatment is that in Treatment 1 the workplace norms announcement makes workers aware of the firm's expectations of worker's (high) effort.

(3) Treatment 2 (Social norms and shame-aversion)

The only difference than baseline treatment is that in Treatment 2 the workers are given information that their social group expect they ought to choose a `high' effort level. Subjects also receive information on the effort level chosen by other members of the social group in similar experiments. If a worker falls short of the effort expectations of their social group, they could be sanctioned by the social group. Such sanctions take the form of social disapproval of `low' effort that falls below the expectations of the social group.

**2.2. Model**

A firm (principal) hires a worker (agent) to work on a project. The worker chooses an effort level $e \in \{e_L, e_H\}$ such that $0 \leq e_L < e_H$ and the respective cost of effort is $0 \leq c_L < c_H$. The chosen effort level is obervable only to the worker, but not to the firm. Alternatively, the firm cannot produce a verifiable signal of effort to a third party.

The firm's production technology is stochastic, giving rise to two possible output levels $0 \leq \pi_0 < \pi_1$. The worker's chosen effort level induces the following conditional probability distribution $f: \{\pi_0, \pi_1\} \times \{e_L, e_H\} \to [0,1]$ over the output levels.

$$f(\pi_0|e_L) > f(\pi_0|e_H);\ f(\pi_1|e_L) < f(\pi_1|e_H)$$

Let $f(\pi_1|e_L) = p;\ f(\pi_1|e_H) = q$. Then $f(\pi_0|e_L) = 1 - p, f(\pi_0|e_H) = 1 - q$, and $p < q$.

The firm cannot offer an effort-contingent contract to the worker because effort is unobservable, or unverifiable to a third party. But the firm's output is mutually observed by both parties and verifiable to a third party. Hence, the firm offers the following contract to the worker: If the output is $\pi_1$ the worker receives the wage $w_1$ and if the output is $\pi_0$, the worker receives the wage $w_0$.

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

$$U_L = (1-p)u(w_0) + pu(w_1) - c_L; if\ e = e_L.$$
$$U_H = (1-q)u(w_0) + qu(w_1) - c_H; if\ e = e_H.$$

The firm wishes to maximize expected profits, given by

$$E\pi_L = (1-p)(\pi_0 - w_0) + p(\pi_1 - w_1); if\ e = e_L.$$
$$E\pi_H = (1-q)(\pi_0 - w_0) + q(\pi_1 - w_1); if\ e = e_H.$$

We assume that the firm always likes the worker to put in a high effort level $e_H$ and we are mainly interested in the problem of implementing $e_H$. We require $E\pi_L < E\pi_H$.

Define the following differences which allow for a more compact notation.

$$\Delta u = u(w_1) - u(w_0);\ \Delta\pi = \pi_1 - \pi_0;\ \Delta w = w_1 - w_0;\ \Delta c = c_H - c_L.$$

This requires $\Delta w < \Delta\pi$.

We are not, however, interested in this paper in the design of optimal contracts.

**2.3. Predictions of the theoretical models**

**2.3.1. Treatment 0: The classical principal-agent analysis**

In neoclassical principal-agent models, the incentive compatibility condition (ICC) requires that to choose $e_H$ over $e_L$ we must have

$$ICC: U_H > U_L \Leftrightarrow (q-p)\Delta u > \Delta c.$$

The individual rationality constraint from exerting an effort $e_H$, is satisfied if $U_H \geq 0$, or

$$IRC: u(w_0) + q\Delta u - c_H \geq 0.$$

We shall be interested in situations where the ICC is violated, yet on account of psychological or social factors, workers choose $e_H$, i.e.,

$$\Delta u < \frac{\Delta c}{q-p}.$$

Since lab experiments are typically small stake experiments, and if subjects are approximately risk neutral over small stakes, then $\Delta u \approx \Delta w$ (Rabin, 2000). In this case, we can rewrite the violation of the ICC as

$$w_1 < w_0 + \frac{\Delta c}{q-p} \equiv a_1.$$

Our results are robust to considerations of risk aversion.

**2.3.2. Treatment 1: Guilt aversion in principal-agent contracts**

Treatment 1 incorporates the effects of firm culture or corporate norms, that operate through the channel of guilt-aversion, into the classical principal-agent problem. A guilt-averse worker might feel guilty about putting in an effort level below the expectations of the firm, $e_H$, which in turn reflect the firm culture. We use the term guilt-aversion in the sense in which Battigalli and Dufwenberg (2007) use "simple guilt."

The Stage 0 announcement of $e_H$ makes explicit the (first order) beliefs of the firm to the worker. Hence, the worker can accurately form (second order) beliefs about the expectations of the firm. The worker suffers a guilt cost that is increasing in the shortfall in performance relative to the expectations of the firm.

This is captured by a guilt aversion function

$$g = \max\{\Delta e_j, 0\}; j = L, H.$$

where $\Delta e_j = e_H - e_j$.

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

$$V_L = (1-p)u(w_0) + pu(w_1) - c_L - \lambda \Delta e_L,$$

where $\lambda \epsilon [0,1]$ is the relative weight put on guilt aversion.

The incentive compatibility condition for the worker is

$$ICC: V_H > V_L \Leftrightarrow (q-p)\Delta u > \Delta c - \lambda \Delta e_L.$$

The individual rationality constraint from exerting $e_H$, under guilt aversion, is identical to that in T0, because $V_H \equiv U_H$.

In the presence of guilt aversion, the condition for the violation of the ICC is

$$w_1 < w_0 + \frac{\Delta c - \lambda \Delta e_L}{q-p} \equiv a_2.$$

Predictions in Treatment 1:

(1) A larger percentage of subjects to choose the high effort level in Treatment 1 as compared to Treatment 0.

(2) An increase in the size of $\Delta e_L$ should produce an increase in the percentage of subjects who choose the higher effort level in Treatment 1. By contrast there should be no effect on effort of a variation in $\Delta e_L$ in the classical principal-agent model (T0).

The qualitative results above are unchanged if we assume that the decision maker is risk averse.

**2.3.3. Shame aversion in principal-agent contracts**

Treatment 2 incorporates the role of social norms, underpinned by shame-aversion into the classical principal-agent problem. The emotion of shame arises from falling below the expectations of one's social or peer group.

Successful social norms require the satisfaction of three key conditions: Empirical expectations, normative expectations, and sanctions. Successful social norms typically require consistency between empirical and normative expectations. In our experimental design, we ensure that the relevant conditions for social norms 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 worker suffers a potential disutility cost on account of shame-aversion that is increasing in the shortfall in effort relative to the expectations of the social/peer group. This is captured by a shame-aversion function s

$$s = \max\{\Delta e_j, 0\}; j = L, H.$$

where $\Delta e_j = e_H - e_j$.

Denote the utility of the worker from effort level $e_j$ by $W_j$ . In the presence of shame-aversion, the worker's utility function from choosing the effort level $e_L$, is given by

$$W_L = (1-p)u(w_0) + pu(w_1) - c_L - \mu(x)\Delta e_L,$$

where $\mu(x) \epsilon [0,1]$ is the relative weight put on shame-aversion. This weight is increasing in the percentage of other social group members, $x$, who expect group members to exert the high effort level, $e_H$, thus we have

$$\mu' > 0.$$

It follows that the stronger is the normative injunction, the higher the shame-aversion that is felt by the worker from violating the norm.

The incentive compatibility condition for the worker is

$$ICC: W_H > W_L \Leftrightarrow (q-p)\Delta u > \Delta c - \mu(x)\Delta e_L.$$

Suppose that the worker is risk neutral over small stakes. In the presence of shame aversion, the condition for the violation of the ICC is

$$w_1 < w_0 + \frac{\Delta c - \mu(x)\Delta e_L}{q-p} \equiv a_3.$$

Predictions in Treatment 2:

(1) A larger percentage of subjects to choose the high effort level in Treatment 2 as compared to Treatment 0.

(2) An increase in the size of $\Delta e_L$ and $x$ should produce an increase in the percentage of subjects who choose the higher effort level in Treatment 2.

The qualitative results above are unchanged if we assume that the decision maker is risk averse.

We cannot, however, predict which of the two cases, guilt aversion, or shame aversion will produce a higher percentage of subjects who choose the higher effort level. This is an empirical question. In the interval $a_2 < w_1 < a_1$, subjects will choose the higher effort level under guilt-aversion. In the interval $a_3 < w_1 < a_1$, subjects will choose the higher effort level under shame-aversion. Thus, the relative efficacy of the two effects depends on the relative sizes of $a_2$ and $a_3$:

$$a_2 \gtreqless a_3 \Leftrightarrow \lambda \lesseqgtr \mu(x).$$

Thus, 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 $a_2 > a_3$. In other words, for a larger wage interval $a_3 < w_1 < a_1$ it is the case that subjects will choose the higher effort level. Thus, stronger social norms of higher effort are more likely to produce higher effort relative to a reliance on guilt aversion alone.

**2.3.4. Optimal effort under fixed wages**

Classical principal-agent theory also usefully identifies the tradeoff between insurance and incentives. The main insight is that a fixed wage provides insurance but no incentives to choose $e_H$ over $e_L$. We also wish to test this result. Suppose that the wage is constant, irrespective of the output realization, so that $w_1 = w_0 = \widehat{w}$. In particular, we choose $\widehat{w} = (1-q)w_0 + qw_1$, so the expected profits of the firm are not affected by paying a fixed wage.

Under fixed wages, in classical principal-agent models, the worker is fully insured but has no incentives to exert any effort. However, the worker might be motivated by a firm culture of high effort level and choose the high effort if guilt aversion is high enough. In the presence of social norms of high effort, the worker might also choose the high effort if shame-aversion is high enough. If we observe workers in Treatments 1 and 2 to put in high effort level under fixed wages, then we can derive lower limits on their guilt-aversion and shame-aversion parameters.

## 2.4. Parameterization of the experiments

### 2.4.1 Parameters and rationale for the choice of parameters

1. We choose the contractual parameters $w_0, w_1$ such that two conditions jointly hold.

The firm finds it optimal to choose the high effort level, so that $\Delta w < \Delta\pi$. This is to ensure that the contractual parameters chosen by the firm are plausible. Principal-agent theory requires the output-state, $\pi_0$ (bad state) or $\pi_1$ (good stage) to be verifiable to both parties, as well as the ex-ante probability distribution over them when the worker chooses each effort level. However, while the exact numbers $\pi_0, \pi_1$ need to be known to the firm, they need not be known to the worker. We design a procedure below such that both parties, the principal and the agent, are able to view the realization of the output-state, but only the firm observes the actual numbers $\pi_0, \pi_1$ in the respective states. This design is necessary to minimize the role of social comparisons and inequity aversion considerations in our analysis. We are not arguing that such comparisons might not be important in the real world if workers can indeed observe the state-wise profits of firms, however, we wish to engage in as clean a test of our hypotheses as possible.

2. The ICC is violated in the classical principal-agent problem, so that under risk neutrality holds, i.e., $\Delta w < \frac{\Delta c}{q-p}$.

Under these two conditions, in the classical principal-agent model, the worker should not choose $e_H$. We choose the following numbers for the experiment, expressed in terms of tokens (1 token=0.15 Yuan). For our first set of parameters, we choose the following parameters.

$$w_0 = 200, w_1 = 400; \pi_0 = 600, \pi_1 = 850.$$

This ensures that the condition $\Delta w = 200 < \Delta\pi = 250$ holds. Thus, the firm always finds it beneficial to offer the given contract.

For the second set of parameters we assume that

$$e_L = 7{,}5{,}3; e_H = 8;\ c_L = 100;\ c_H = 200; p = 0.3, q = 0.7.$$

Notice that we try 3 different values of $e_L$ because we wish to employ the strategy method to vary $\Delta e_L$. Hence, we have $\Delta c = 100$, $q - p = 0.4$. Thus, the parameters satisfy $\Delta w = 200 < \frac{\Delta c}{q-p} = \frac{100}{0.4} = 250$.

**2.4.2 Predictions for the case of monotonically increasing wages**

1. In Treatment 1, in the presence of firm culture and guilt-aversion, the ICC is satisfied if $w_1 > w_0 + \frac{\Delta c - \lambda \Delta e_L}{q-p}$, or equivalently if

$$\lambda > \frac{\Delta c - \Delta w(q-p)}{\Delta e_L} = 0.4, e_L = 3$$

In other words, if the worker puts at least 40% weight on guilt-aversion relative to the maximization of material payoffs, then the guilt-averse worker will choose to exert the high effort level, even when the classical principal-agent model predicts the choice of the low effort level.

For the other two values of $e_L = 7{,}5$ we have $\lambda > 2, e_L = 7;\ \lambda > 0.67, e_L = 5$.

2. In Treatment 2, in the presence of social norms and shame-aversion, the ICC is satisfied if $w_1 > w_0 + \frac{\Delta c - \mu(x) \Delta e_L}{q-p}$, or equivalently if

$$\mu > \frac{\Delta c - \Delta w(q-p)}{\Delta e_L} = 0.4$$

In other words, if the worker puts at least 40% weight on shame-aversion relative to the maximization of material payoffs, then the shame-averse worker will choose to exert the high effort level, even when the classical principal-agent model predicts the choice of the low effort level.

For the other two values of $e_L = 7{,}5$ we have $\lambda > 2, e_L = 7;\ \lambda > 0.67, e_L = 5$.

**2.4.3 Predictions for the case of fixed wages**

Consider $e_L = 3$ (the other 2 cases can be constructed analogously). The worker finds it optimal to choose $e_L$ in the classical model in Treatment 0. We have $\frac{\Delta c}{\Delta e_L} = \frac{10}{5} = 2$. In the presence of firm culture in Treatment 1, the worker chooses $e_H$ if $\frac{\Delta c}{\Delta e_L} = 2 < \lambda$. In the presence of social norms in Treatment 2, the worker chooses $e_H$ if $\frac{\Delta c}{\Delta e_L} = 2 < \mu$.

Thus, the conditions for choosing the high effort under fixed wages is more stringent relative to monotonically increasing wages. We now require the worker to place twice as much weight on guilt-aversion parameter relative to material payoff in Treatment 1; and twice as much weight on shame-aversion, relative to material payoffs in Treatment 2. It is also possible that the choice of fixed wages, instead of incentive-based wages, conveys to the worker a kindness intention on the part of the firm (Brown et al., 2004; Fehr et al., 2007). Hence, workers respond differently to fixed wages and feel more guilt-averse and/or shame-averse from letting down expectations of high effort levels. Whether a greater/lower proportion of workers chooses the high effort level under fixed wages is an empirical question that cannot be answered on theoretical grounds.

**2.5. Data Collection**

We will conduct the experiment in lab in a Chinese university. Research Assistants with the necessary human subjects research training and prior experience with administering lab experiments will be trained to administer this experiment.

## 3. Empirical Analysis

### 3.1. Variables

For each individual, we elicit their effort levels in each treatment. Table 1 describes outcome variables.

Table 1: Primary Outcome Measures

| Outcome Variable | Description | Outcome Measures | Outcome Measure Type |
|---|---|---|---|
| **T0** | | | |
| $F_1$ | Firms are asked to choose whether to offer the fixed wage contract | Minimum = 0<br>Maximum = 1 | Binary variable<br>{0,1} |
| $F_2$ | Firms are asked to choose whether to offer the varied wage contract | Minimum = 0<br>Maximum = 1 | Binary variable<br>{0,1} |
| $W_1$ | Workers are asked to choose whether to refuse the fixed wage contract | Minimum = 0<br>Maximum = 1 | Binary variable<br>{0,1} |
| $W_2$ | Workers are asked to choose whether to refuse the varied wage contract | Minimum = 0<br>Maximum = 1 | Binary variable<br>{0,1} |
| $W_3$ | Workers are asked to choose their preferred effort levels (low effort or high effort) under different low effort levels in both fixed and varied wage cases (with no information about anyone's expectations) | Minimum= 0<br>Maximum= 1 | Binary variable<br>{0,1} |
| **T1 (in addition to the first 4 variables in T0)** | | | |
| $W_4$ | Workers are asked to choose their preferred effort levels (low effort or high effort) under different low effort levels in both fixed and varied wage cases (with the information that the firm expects high effort) | Minimum= 0<br>Maximum= 1 | Binary variable<br>{0,1} |

| | | | |
|---|---|---|---|
| **T2 (in addition to the first 4 variables in T0)** | | | |
| $W_5$ | Workers are asked to choose their preferred effort levels (low effort or high effort) under different low effort levels in both fixed and varied wage cases, with the information of (1) different percentages of social group believe workers ought to exert high effort, and (2) the percentage of social group chose high effort in previous similar experiment | Minimum= 0<br>Maximum= 1 | Binary variable<br>{0,1} |

Table 2: Secondary outcome measures

| Outcome Variable | Description | Outcome Measures | Outcome Measure Type |
|---|---|---|---|
| **T1** | | | |
| Guilt | Workers are asked whether they would feel guilt from falling behind the firm's expectations of high effort | Minimum = 1 ("I feel no guilt")<br>Maximum = 10 ("I feel maximum possible guilt") | Ordinal variable<br>[1, 10] |
| **T2** | | | |
| Shame | Workers are asked whether they would feel shame from falling behind the expectations of their social group | Minimum = 1 ("I feel no shame")<br>Maximum = 10 ("I feel maximum possible shame") | Ordinal variable<br>[1, 10] |
| **T0 and T1** | | | |
| Disapproval | Workers are asked whether disapproving from the low effort choices from their social group | Yes/No | Binary variable<br>[0,1] |
| NE | Workers are asked whether they believe their social group ought to exert high effort | Yes/No | Binary variable<br>[0,1] |

The following variables will be used as covariates in regression analysis.

Table 3: Covariates

| **Type** | **Description** | **Measure creation/ Final variable used** |
|---|---|---|
| **Covariates:** Demographics | Age of respondent | No manipulation needed |
| | Gender of respondent | Nominal variable |
| | Field of study | Binary variable |
| | Similar experiment experience | Binary variable |
| | Annual household income | Categorical variable |

### 3.2. Model Specifications

In addition to some descriptive analysis, we can run regressions using our primary outcome measures and our treatment effects with and without our control variables.

In details, we can use Probit models to estimate the effects which influence the worker's choice of high or low efforts. The dependent variable is the probability of the worker's choices of high effort ($P(High)$), and the independent variables can include the signal of the signal of the normative expectation (NE), if firm's high effort expectation exists (firm), fixed wage or variable wage (fixed), $\Delta e_L$, and the demographic variables (X).

For example, using the data of T0 and T2, we can run the following model to mainly test the predictions listed in Treatment 2.

$$P(High) = \alpha_0 + \alpha_1 NE + \alpha_2 fixed + \alpha_3 \Delta e_L + \alpha_4 X + u$$

According to our model and the literature, we predict that the estimation of $\alpha_1$ is positive, and the estimation of $\alpha_3$ is positive as well.

While using the data of T0 and T1, we can run the following model to mainly test the predictions listed in Treatment 1.

$$P(High) = \beta_0 + \beta_1 firm + \beta_2 fixed + \beta_3 \Delta e_L + \beta_4 X + u'$$

According to our model, we predict that the estimation of $\beta_1$ and $\beta_3$ are positive.

Besides, if the signs of $\beta_2$ and $\alpha_2$ are positive is to be observed from the data.

We will also consider some interacting effects between variables in addition to the above models.