THREE ESSAYS ON SUBJECTIVE PERFORMANCE EVALUATION

QIAN NENG

(B.A., FUDAN UNIVERSITY, 2006;

M.A., FUDAN UNIVERISITY, 2009)

A THESIS SUBMITTED FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECONOMICS

NATIONAL UNIVERSITY OF SINGAPORE

May, 2014

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

钱能

QIAN NENG 28 May 2014

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my supervisor, Professor Parimal Bag, who has the attitude and substance of a real scholar. All three chapters of my thesis are joint work with him, and the thesis would have remained a dream without his guidance and encouragement. I would also like to thank my committee members, Professor Jingfeng Lu, Professor Julian Wright and Professor Qiang Fu for their support throughout my candidature. Special thanks should be given to Professor Satoru Takahashi, who has given valuable comments and suggestions to my work. I also owe my gratitude to Professors Xiao Luo, Yi-Chun Chen, Chiu Yu Ko and others for their sincere support and help in different stages.

In addition, I would like to acknowledge my colleagues, Li Jingping, Liu Zhengning, Lu Yunfeng, Wang Ben, Wang Peng, and others, for the spirit and belief among us - we are the best! Last but not least, thanks to my family for your trust and love. I do appreciate the unique experience of the five years in my life.

Contents

SUMMARY

This dissertation contains three chapters on the contracting problem under subjective performance evaluation. The first two chapters mainly deal with the money burning contract in a single agent model, complementing the existing literature in understanding the optimal contract form under subjective performance evaluation. The third chapter extends the work into a multi-agent model, investigating the implications of subjective performance evaluation and money burning in a team environment.

In chapter one, I review the work of William Fuchs (2007, *AER*), who proposes that to implement that an agent exerts effort in every period of a finitely repeated 0-1 effort choice game, the principal should penalize the agent by money burning only when he observes low-performance signals in every round. While he is minimizing the expected money burning, we show that Fuchs' mechanism also often maximizes the up-front payment that the principal has to incur for his objective. This dichotomy arises because minimizing expected money burning is not necessarily the dual of the principal's profit maximization problem. For the latter, the principal is better off to rely, most of the time, on disciplining the agent by burning money at even the slightest hint of shirking in any round and increase it with more and more evidence of shirking. In law and economics, this mechanism is known as *penalty fitting the crime*. Also it is shown that the principal is (weakly) better off not to carry out interim performance evaluation or engage in interim money burning. These results are derived in a two-period game.

In chapter two, I further investigate the more fundamental problem in the literature on subjective evaluation: the result of wage compression, based on the work by MacLeod (2003, *AER*), in addition to the previous observation on Fuchs (2007). Optimal effort incentives in contracting under subjective evaluation recommend

that the principal should burn money to slash rewards only when the agent's performance is at its worst possible, but otherwise there should be no penalty and the rewards should be uniform. This *extreme wage compression* hypothesis has been established in two alternative formulations: (i) a static model of a profit-maximizing principal dealing with a risk-averse agent whose utility of money is unbounded from below (MacLeod, 2003), (ii) a finitely repeated game with risk-neutral agent but the principal pursues a social efficiency objective (Fuchs, 2007). Modifying the principal's objective from social efficiency to profit maximization in Fuchs' model, and in MacLeod's model by allowing for more general risk preferences (including risk neutrality) and dropping the assumption of ruin (negative unbounded utility/Inada condition at zero consumption), the optimal contract is shown to be either of the *pay for performance* type where rewards gradually improve with performance (Holmstrom, 1979; Harris and Raviv, 1979), or one of *moderate wage compression* with zero reward below a threshold performance and full reward above the threshold, similar to Levin's (2003) *termination contract*. The extreme wage compression result with money burning (or penalty) restricted to a single low incidence, worst performance signal is thus a special case of more general possibilities.

In chapter three, I study the optimal contracting problem under subjective performance evaluation in teams. We find that, absent verifiability the principal relies on subjective evaluation of team performance and must burn money for poor performance, which can be interpreted as passing on the rewards to non-critical employees. Such *"must spend"* mechanisms along with discriminatory treatment of agents tend to create a culture of sabotage that it might not be possible for the principal to prevent. And even when sabotage can be deterred, its very possibility may increase the costs of implementing full team efforts. Ultimately, the power of subjective performance evaluation gets eroded due to back-stabbing and scheming within teams. Given that money burning, or blatantly wasteful spending, is not really a choice for most organizations, one might be left with only a *scheming* group. This is in addition to the familiar problem of collusion encountered in team settings.

List of Figures

List of Tables

CHAPTER 1

Revisiting Money Burning in Performance Evaluation

1.1 Introduction

Asking an agent to perform a task repeatedly, exert effort or shirk, when the principal privately observes (signals of) the agent's performance but not effort choices is a natural extension of the issue of *subjective performance evaluation*, earlier studied by Bentley W. MacLeod (2003) and Jonathan Levin (2003). William Fuchs (2007) analyzes this incentive provision problem with a number of interesting observations on: (i) how the agent should be rewarded or penalized over time as the principal closely follows the agent's track record, (ii) should the agent be given real time feedback, (iii) how sensitive the rewards should be to the intertemporal structure of performance, etc. Among these, one particular observation is quite striking. The author notes that in the finitely repeated game the principal should penalize the agent by burning money only when the agent's performance exhibits the extremely unlikely sequence of *all* low signals. Taken literally, when $T = 10$ the principal penalizes the agent when $(\underbrace{\sigma_L \sigma_L ... \sigma_L}_{10 \text{ times}})$ materializes whereas in all other $2^{10} - 1$ other sequences of signals involving at least one σ_H, the agent is completely let off the hook, i.e., no money will be burnt to slash a prior committed exogenous reward of w. See Proposition 3 in Fuchs' article.

In this paper, we review the above recommendation of Fuchs. The observation is striking because most organizations, we believe, are unlikely to adopt such a 'generous' approach to incentivize their employees. One may also want to be careful in labelling the incentive as generous because when all low signals do materialize, the money burnt will be substantial. That is, the penalty is huge. But then the principal-committed rewards must be very large too unless, of course, the agent puts up a significant amount in bond before agreeing to work for his employer. Let's say most employment situations do not require such bonds. Then the principal must fork out the big rewards mainly to threaten the agent to blow it up when all signals stack up against the agent. We are going to argue that if the objective of the incentive mechanism is to minimize the *principal's implementation costs* of inducing the agent to exert efforts in all of T rounds, the principal ought to penalize the agent at the slightest hint of shirking and follow the standard law-and-economics doctrine of *penalty fitting the crime* (see, for instance, James Andreoni (1991), or Steven Shavell (1991)). Our suggested mechanism is noteworthy as it contrasts sharply with the one in Fuchs (2007). And in terms of description of organizational behavior ours is perhaps closer to the actual practice than one prescribed by Fuchs, although we do not claim to take a definite stand on this.

At this stage we should note that Fuchs' result derives from an entirely different premise, maximization of the objective of *social optimality* rather than minimization of principal's costs. Social optimality requires minimization of expected money burning, as it is a deadweight loss, subject to implementation of agent efforts in all T rounds. But if one wants to understand organizational behavior, the relevant objective should be the principal's profit maximization or, equivalently for any sequence of efforts, cost minimization, which under appropriate assumption amounts to minimizing the maximal amount of money burning needed over all sequences of signal realizations for the targeted efforts. Only in some situations with rather weak associativity between effort and high performance signal, the two objectives will yield identical penalties (or money burning). In this paper, we complement Fuchs (2007) by shifting the focus from social optimality to principal's cost minimization.

Besides suggesting that the agent's penalty based on performance should be of the more conventional type, we also show that for the modified objective the principal should provide no interim feedback just like what Fuchs has argued. Instead, the principal should wait till the end and burn money in proportion to the evidence of low performance in all the rounds combined.

The rest of the paper is organized as follows. In the next section we present the model. Our main analysis and the results are contained in sections 3 and 4. We close with some final remarks in section 5. Proofs appear in the Appendix.

1.2 The model

A risk-neutral principal involves a risk-neutral agent in a T-period repeated efforts game. In each period the agent can either exert one unit of effort or shirk, $e_t \in \{0,1\}$, with effort costing the agent $c > 0$. There is no discounting by the principal or the agent. The principal does not observe the agent's effort choice but receives a private signal $\sigma_t \in \{\sigma_H, \sigma_L\}$ of agent performance that cannot be disclosed verifiably. During a particular round the performance signal depends only on the effort in that round as follows:

$$\Pr[\sigma_t = \sigma_H \mid e_t = 0] = p_0, \qquad \Pr[\sigma_t = \sigma_H \mid e_t = 1] = p_1,$$

with $0 < p_0 < p_1 < 1$.

At the end of T periods, the principal will report the performance signals $S^T \in \Sigma = \{\sigma_1\sigma_2...\sigma_T\}$. Let $w^P(S^T)$ be the amount of money pre-committed to be *spent* by the principal for any reported profile of performance signals. Due to the subjective performance evaluation, the principal has to pay a fixed amount regardless of the signal profile in order for him not to misreport the signals. However, to incentivize the agent to exert effort, the reward should be contingent on the reported performance S^T. Thus, the agent should not receive all the money paid by the principal under some circumstances so that budget balance may break down, which formally amounts to *money burning*.[1] Denote the fixed budget for the principal as $w^P(S^T) = W$, the contingent reward for the agent as $w^A(S^T)$, and the amount of money burning as $z(S^T)$. We use bold symbols $\mathbf{w^A}(\Sigma)$ and $\mathbf{z}(\Sigma)$ to denote vectors of rewards and money burning corresponding to signal profiles $S^T \in \Sigma$. Formally, the contract is defined as:

$$\begin{aligned} \omega &\equiv \left(W, \mathbf{z}(\Sigma), \mathbf{w^A}(\Sigma)\right), \quad \text{where} \\ W &= w^A(S^T) + z(S^T) \qquad \forall S^T. \end{aligned} \tag{1.1}$$

Given the above, we will write simply $\omega \equiv (W, \mathbf{z}(\Sigma))$ to refer to the incentive mechanism.

Let $\mathbf{e} = (e_1, \cdots, e_T)$ be the agent's efforts over T rounds. Denote the probability of signal profile S^T conditional on $\mathbf{e}$ by $P(S^T \mid \mathbf{e})$. Given the incentives, the agent's expected reward can be written as:

$$\begin{aligned} V(\mathbf{e}) &= \mathbb{E}\left(w^A(S^T) \mid \mathbf{e}\right) - c\sum_t e_t \\ &= W - \sum_{S^T} z(S^T)\, P(S^T \mid \mathbf{e}) - c\sum_t e_t. \end{aligned} \tag{1.2}$$

[1] See MacLeod (2003), who proposed the idea of money burning to solve the problem of subjective performance evaluation. Fuchs (2007) also shows a similar result.

We assume that no interim feedback or payment/money burning is allowed.[2] The time line of the game is as follows:

1. *At time zero, a contract* $(W, \mathbf{z}(\Sigma))$ *is signed between parties;*

2. *At each period* $t = 1, 2, ..., T$*, the agent decides whether to exert effort on the principal's project or shirk and the principal receives a performance signal at the end of the period;*

3. *At the end of* T *periods, the principal reports the performance signal profile* S^T *and makes the payment* $w^A(S^T)$*.*

The incremental expected value of output in any period following change from $e = 0$ to $e = 1$ in that round is assumed to be large enough so that the principal wants to design an incentive compatible rewards scheme $(W, \mathbf{z}(\Sigma))$ to uniquely implement $\mathbf{e}^* = (1, 1, ..., 1)$ at minimal W.[3] Formally, the principal solves the following problem:

$$\min_{W, \mathbf{z}(\Sigma)} \quad W \tag{$\mathcal{P}_1$}$$

$$\text{s.t.} \quad {\scriptstyle \text{(Incentive Compatibility)}} \quad \mathbf{e}^* \in \arg\max V(\mathbf{e}), \tag{1.3}$$

$$\quad {\scriptstyle \text{(Limited Liability)}} \quad W - z(S^T) \geq 0 \quad \forall S^T, \tag{1.4}$$

$$z(S^T) \geq 0 \quad \forall S^T. \tag{1.5}$$

Incentive compatibility (1.3) and limited liability (1.4) will guarantee agent's participation constraint $V(\mathbf{e}^*) \geq 0$. Let us denote by $\omega^\star := (W^\star, \mathbf{z}^\star(\Sigma))$ the optimal money burning contract solving problem ($\mathcal{P}_1$) that implements $\mathbf{e}^*$.

[2] We will relax this assumption later.

[3] Proposition 3 in Fuchs (2007) mentions implementation of agent exerting efforts in *all* periods. In his Lemma 2, Fuchs states that the principal can always offer a payoff-equivalent contract with the agent exerting efforts in every period.

1.3 The optimal mechanism of Fuchs (2007)

Fuchs (2007) studied the same T-rounds effort implementation problem using a money burning contract but with an important difference: instead of principal's reward costs, the author minimized expected money burning. The idea must have been that since money burning is a social loss, minimizing its expected value solves the second-best program.[4] Thus, Fuchs actually solved the social planner's problem and his optimal contract characterization is not necessarily the same as maximizing the principal's selfish surplus maximization objective. It is this latter task that we focus on. In order to highlight the contrast between our optimal mechanism and that of Fuchs, next we report Fuchs' formulation of the problem, the associated mechanism and some of its properties.

Fuchs' principal solves:[5]

$$\min_{\mathbf{z}(\Sigma)} \quad \mathbb{E}\left(z(S^T) \mid \mathbf{e}^*\right) \tag{$\mathcal{P}_2$}$$

$$\text{s.t.} \qquad (1.3) \quad \text{and} \quad (1.5).$$

We can now see the difference between our optimization problem ($\mathcal{P}_1$) and Fuchs' problem ($\mathcal{P}_2$). Take any solution $\mathbf{z}(\Sigma)$ to ($\mathcal{P}_2$) and define $W = \max_{S^T}\{z(S^T)\}$ so that it satisfies, by construction, (1.4); thus solution to problem ($\mathcal{P}_2$) cannot dominate, in terms of principal's implementation costs, the solution to problem ($\mathcal{P}_1$). But it is possible that

[4] Fuchs first defined the optimal contract without money burning from the principal's point of view to be one that maximizes expected discounted value of output net of the wages (see section I), then he looked at the case with money burning (section II). For the latter, his optimal contract exhibits two characteristics: no interim feedback and transfers or money burning until in the last period (Lemma 1), and the agent exerts efforts in every period (Lemma 2). In our formulation, initially we assume no-interim-feedback and incremental per-period output from effort to be high enough to justify implementation of $\mathbf{e}^* = (1, 1, ..., 1)$ as part of the principal's cost-minimizing contract under no discounting that would also maximize his net private surplus. Later on we verify no-interim-feedback to be optimal.

[5] Here we take it as given that the principal would induce $\mathbf{e}^*$.

there is some other $\tilde{\mathbf{z}}(\Sigma)$ satisfying (1.3) and (1.5) that do not solve problem ($\mathcal{P}_2$), yet $\max_{S^T}\{\tilde{z}(S^T)\} < \max_{S^T}\{z(S^T)\}$; such a case is illustrated in Fig. 2.1 in section 5 and will be more explicitly shown in Proposition 2. Now define $\tilde{W} = \max_{S^T}\{\tilde{z}(S^T)\}$ so that $(\tilde{W}, \tilde{\mathbf{z}}(\Sigma))$ satisfy (1.3)-(1.5) and thus strictly dominates $(W, \mathbf{z}(\Sigma))$ for the principal's cost-minimization problem ($\mathcal{P}_1$). In other words,

PROPOSITION 1 (**Failure of Fuchs' mechanism for cost minimization**). *Fuchs' optimal money burning mechanism may fail to achieve the minimal* e^**-implementation costs for the principal and can* <u>*never*</u> *do strictly better than the solution to problem* ($\mathcal{P}_1$)*.*

The above observation does not tell us yet when Fuchs' mechanism might fail in achieving principal's cost-minimization objective. To answer this we need to solve for the optimal money burning mechanism for the problem ($\mathcal{P}_1$), which we address in the next section. Below we study Fuchs' mechanism more closely.

Fuchs' money burning mechanism can be explicitly written as follows (derives from Proposition 3 in Fuchs (2007)):

$$\begin{cases} & w^A = W - Z \quad \text{if } S^T = \sigma_L\sigma_L...\sigma_L \quad (\text{i.e. } \sigma_t = \sigma_L \ \forall t) \\ & w^A = W \quad \text{otherwise,} \\ \text{where} & Z = \dfrac{c}{(p_1 - p_0)(1-p_1)^{T-1}} \\ & W = \max\left\{\dfrac{c}{(p_1 - p_0)(1-p_1)^{T-1}}\,,\, Tc + \dfrac{(1-p_1)c}{(p_1 - p_0)}\right\}. \end{cases}$$

To minimize the expected amount of money burning, the agent's penalty takes a positive value, Z, only when the signals in all T periods are low, and in all other cases with at least one high signal the money burning is zero. The probability that all signals will be low is very small, implying Z must be very high for it to threaten the agent to exert efforts in all the periods. This becomes clear from Fuchs' optimal contract: Z upon worst performance is increasing very fast with T or p_1. Along with this, W is pushed up by the requirement that the agent's reward must be non-negative: $w^A = W - Z \geq 0$. Intuitively, by penalizing the agent with an extremely low frequency, the money burning upon the worst situation and the corresponding budget to cover for the money burning becomes very large. After all, the principal would design incentives with his own costs in mind rather than just saving wasteful money burning in expected terms. The point of our exercise is to clarify this aspect, the contrast between what might be socially optimal and what the organization should prefer. Social optimality, as shown by Fuchs' analysis, recommends an extreme and unlikely penalty prescription. As we will see in the next section, an organization should like to adopt a more routine approach to penalty: find out the number low signals of performance and penalize in an increasing order.

1.4 Cost-minimizing money burning: Two-period case

In this section, we analyze the optimal money burning incentives for the cost-minimization objective of the principal. The basic message can be easily conveyed by studying a two-period effort implementation problem. Initially we proceed under the assumption of no interim money burning, then we discuss its plausibility.

■ No interim money burning. A principal hires an agent to work for two periods without discounting. The agent's strategies are $(e_1, e_2) \in \{(1,1), (1,0), (0,1), (0,0)\}$, and possible signals, S^T, are $\{\sigma_H\sigma_H, \sigma_H\sigma_L, \sigma_L\sigma_H, \sigma_L\sigma_L\}$. The incentives can be written as $\omega = (W, \mathbf{z}(\Sigma))$, where $\mathbf{z}(\Sigma) = \{z^{HH}, z^{HL}, z^{LH}, z^{LL}\}$. The agent's payoffs in the repeated efforts game is illustrated in Fig. 1.1.

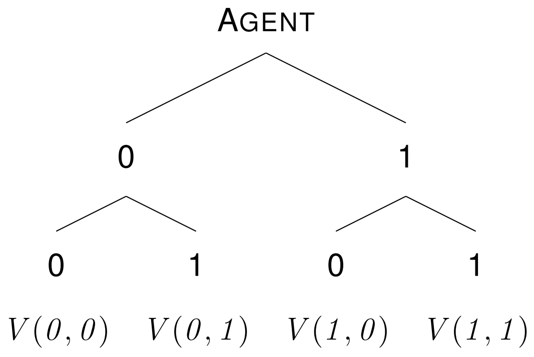

Figure 1.1: Repeated efforts game

The principal wants to induce $(e_1, e_2) = (1, 1)$ at minimal costs:

$$\begin{aligned} &\min_{W,\mathbf{z}(\Sigma)} \quad W \\ s.t. \qquad &(1,1) \in \arg\max V(e_1, e_2), \\ &W - z(S^T) \geq 0, \\ &z(S^T) \geq 0. \end{aligned}$$

Solving the principal's problem, we can characterize the optimal contract as follows:

PROPOSITION 2 (**Optimal money burning contract**).

(I) *If* $p_1 > \frac{1}{2}$ *and* $p_1 + p_0 > 1$*,*

$$W = \tfrac{2c}{(p_1+p_0)(p_1-p_0)}\,,\; z^{HH} = 0\,,\; z^{LH} = z^{HL} = z^{LL} = \tfrac{2c}{(p_1+p_0)(p_1-p_0)}.$$

(II) *a. If* $p_1 > \frac{1}{2}$ *and* $p_1 + p_0 = 1$*,*

$W = \frac{2c}{p_1-p_0}\,,\; z^{HH} = 0\,,\; z^{LL} = \frac{2c}{p_1-p_0}\,,$

$z^{LH}, z^{HL} \in [\frac{c}{p_1-p_0}, \frac{2c}{p_1-p_0}]$ *and* $p_1 z^{HL} \geq p_0 z^{LH} + c\,,\; p_0 z^{HL} \geq p_1 z^{LH} - c\,.$

b. If $p_1 > \frac{1}{2}$ *and* $p_1 + p_0 < 1$*,*

$$W = \tfrac{2c}{p_1-p_0}\,,\; z^{HH} = 0\,,\; z^{LH} = z^{HL} = \tfrac{c}{p_1-p_0}\,,\; z^{LL} = \tfrac{2c}{p_1-p_0}.$$

c. If $p_1 = \frac{1}{2}$*,*

$$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\; z^{HH} = 0\,,\; z^{LH} = z^{HL} \in [0, \frac{c}{p_1-p_0}]\,,\; z^{LL} = \frac{2c}{p_1-p_0}.$$

(III) *If* $p_1 < \frac{1}{2}$,

$$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\; z^{HH} = z^{LH} = z^{HL} = 0\,,\; z^{LL} = \frac{c}{(1-p_1)(p_1-p_0)}.$$

Thus, the optimal money burning scheme depends on the signal generating technology. When $p_1 > \frac{1}{2}$ and $p_1 + p_0 > 1$, the punishment is most severe and extensive: all money will be burnt unless good signal is received for both periods. This technology implies that, either p_1 is quite high or both p_1 and p_0 are fairly high. If p_1 is close to 1 say, exerting efforts will generate high signals almost surely, so a low signal is an indication of shirking rather than bad luck; if both p_1 and p_0 are high, shirking also has a good chance of generating a high performance signal, thus shirking should be deterred by penalizing when at least one low performance signal is realized. When $p_1 > \frac{1}{2}$ and $p_1 + p_0 \leq 1$, it implies that the difference in the probabilities of generating a high signal when the agent exerts effort vs. when he shirks is going to be non-trivial ($p_1 > \frac{1}{2}$ but $p_0 < \frac{1}{2}$), thus the choice of effort or shirking is very likely to be reflected in the signal generated: signals, although imperfect, are informative. In this case, the money burning scheme is also realistic: all money is burnt if both periods see the low signals; partial money is burnt upon a combination of one low and one high signal. That is, penalty is proportional to (or fits) the "crime" – a general dictum of the law and economics literature (Andreoni, 1991; Shavell, 1991). When $p_1 \leq \frac{1}{2}$ so that the high signal is less likely than the low signal with agent exerting effort, i.e. in the case of weak informativeness of high signal, money burning happens only when both periods have low signals, which coincides with Fuchs' social efficiency maximization prescription:

Minimization of expected money burning. *When* $T = 2$*, the expected* *money burning minimization contract of Fuchs (2007) is given by:*

$$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\ z^{HH} = z^{LH} = z^{HL} = 0\,,\ z^{LL} = \frac{c}{(1-p_1)(p_1-p_0)}.$$

In this last case, while high signal does not suggest a strong evidence of agent's effort, low signal on the other hand would imply a high chance that the agent did *not* exert effort: $p_0 < \frac{1}{2}$. That is, rather than the high signal, its *absence* is more indicative. With $p_0 < p_1$, the principal can rely on the signals' informativeness (as monotone likelihood ratio property [Milgrom, 1981] will be satisfied) to determine money burning.

Comparing the two contracts – ours and that of Fuchs – we can see that when $T = 2$, the solution to Fuchs' problem is optimal for *our* principal only if $p_1 \leq \frac{1}{2}$ as illustrated in part **(III)** of Proposition 2. In parts **(I)** and **(II)** when the chance of generating high signal upon effort is high ($p_1 > \frac{1}{2}$), it is necessary for the principal to burn money upon medium signal profile. Otherwise, it results in insufficient incentives: the agent would work only for one period with the plan of generating either signals $\sigma_H\sigma_L$ or $\sigma_L\sigma_H$. However, when the probability of generating the high signal is not very high even if effort is exerted, money burning does not happen for medium signals since maximum efforts may also lead to this situation. Therefore, any punishment upon medium signal profile would be damaging for effort incentives. In this case, the problem of minimizing expected money burning coincides with *our* principal's problem.

It is also clear that Fuchs' socially efficient contract often results in a "maximal" budget for the principal. That is, whenever our optimal contract differs from Fuchs' optimal contract (as in parts **(I)** and **(II)** of Proposition 2), the following two hold:

1. Maximum money burning across all signal profiles in Fuchs' contract

$>$ maximum money burning in our setting;

2. Expected money burning in Fuchs' contract $<$ expected money burning in our setup.

Observation [1] above was already hinted at in Proposition 1.

■ Interim money burning. We now address the question of interim performance evaluation.[6] With that in mind, consider interim money burning in the two-period game. Suppose after period one the principal announces the realized signal, and carries out the corresponding first-period money burning $z_1 \in \{z_1^H, z_1^L\}$; after period two, the second signal is reported and the follow-up money burning $z_2 \in \{z_2^{HH}, z_2^{HL}, z_2^{LH}, z_2^{LL}\}$ takes place. Now the agent's incentives can be structured in a more piecemeal manner targeted towards each period's effort separately. Is it any better than trying to control two individual efforts with one penalty instrument? In Proposition 4 below we answer this in the negative, but first we report the optimal incentives under interim performance evaluation.

PROPOSITION 3 (**Optimal contract with interim money burning**). *For full efforts implementation* $\mathbf{e}^* = (1,1)$*, the interim money burning contract that minimizes principal's budget is as follows:*

$$
\begin{aligned}
&W = \frac{2c}{p_1 - p_0} \\
&z_1^H = 0\,,\ z_2^{HH} = 0\,,\ z_2^{HL} = \frac{c}{p_1 - p_0} \\
&z_1^L \in [0, \frac{c}{p_1 - p_0}]\,,\ z_2^{LH} = \frac{c}{p_1 - p_0} - z_1^L\,,\ z_2^{LL} = \frac{2c}{p_1 - p_0} - z_1^L.
\end{aligned}
$$

It is reasonable not to punish the agent if first period's or both periods' performance is good ($z_1^H = 0, z_2^{HH} = 0$). However, if the first-period

[6] Issues of interim performance evaluation are starting to gain recognition in formal models with the works of Alessandro Lizzeri, Margaret Meyer and Nicola Persico (2003), Alex Gershkov and Motty Perry (2009), and Masaki Aoyagi (2010), among others.

performance signal turns out to be bad, the optimal contract shows that the principal need not burn interim money: z_1^L can take the value of 0. It is the total money burning that matters to the principal ($z_1^L + z_2^{LH} = \frac{c}{p_1 - p_0}$ and $z_1^L + z_2^{LL} = \frac{2c}{p_1 - p_0}$). Whatever amount of money is burnt in the first period, the principal will 'top up' the penalization in the second period to the targeted total amount. This observation tells us that *interim feedback* (only information communication without actual actions) and actual *interim money burning* have the same effect. This is intuitive as the agent can account for the expected wage reduction at the end of period 1 following an interim report of a bad performance, when it is not followed by immediate (or interim) money burning.

Next, we show that sometimes interim money burning can actually be an inefficient arrangement from principal's point of view:

PROPOSITION 4 (**Sub-optimality of interim money burning**). *For full efforts implementation* $\mathrm{e}^* = (1, 1)$*, interim money burning contract takes the same form as the optimal contract without interim money burning if* $p_1 \geq \frac{1}{2}$ *and* $p_1 + p_0 \leq 1$*; otherwise, it leads to a higher budget for the principal.*

One way to understand why interim feedback and the associated money burning may increase principal's costs is to go back to the incentives in Proposition 2. There, money burning without interim feedback sometimes prescribed either *no money burning* if at least one of the two signals is high (part **III**), or *burning all money* if at least one of the two signals is low (part **I**). With this extreme penalty structure, if interim feedback is introduced then keeping the total amount of money burning unchanged over the same two-period signal profile(s) will damage the agent's first- and/or second-period effort incentives as follows. Consider case (**I**) and let us recall our observation following Proposition 3 that interim money burning is equivalent to interim feedback with only one-time money burning in the end. Now let us see what happens if we were to take the incentives of Proposition 2 and apply it after engaging in only interim feedback. This will clearly destroy the agent's second-period effort incentive following *low* realization of the first-period signal, as whether the second-period signal is low or high the money burning will be the same: the entire reward is to be blown off. This means to restore the agent's second-period effort incentive, we must reset a new money burning pair $(z_2^{LL}, z_2^{LH}) \neq (z^{LL}, z^{LH})$ such that $z_2^{LL} > z_2^{LH}$. This re-configuration will be costly for the principal as he faces more incentive hurdle. In the case of (**III**), both the first and second period incentives get harmed with interim feedback: $z_1^H + z_2^{HL}$ must be strictly positive because z_2^{HL} must be strictly positive as otherwise there is no incentive to exert effort in the second period; but then $z_1^L + z_2^{LL}$ must rise above z^{LL} because $[z_1^L + z_2^{LL}] - [z_1^H + z_2^{HL}]$ provides the incentive for first-period effort and now $z_1^H + z_2^{HL}$ is strictly positive; if $z_1^L + z_2^{LL}$ were at the same level as z^{LL}, the agent's first-period effort incentive would have been weakened and thus failed (since in the original solution for Proposition 2, the agent's first-period effort incentive constraint $V(1,1) - V(0,1) \geq 0$ will be binding; for details refer the proof). This means principal's implementation costs would increase.

As an alternative explanation, we can say that the principal providing the agent with an extra bit of information in the interim period (whether his performance signal is low or high) while keeping the two-period terminal payoffs following each signal profile unchanged can only improve the agent's situation and definitely not worsen relative to when no such interim feedback is provided. This means such information communication will make the principal's incentive provision problem harder and thus more costly.

Under subjective performance evaluation the principal not wanting to carry out interim performance evaluation is puzzling. Many aspects of job evaluations in real life involve subjective assessments by supervisors or managers. Most organizations are also likely to have human resource departments that carry out annual reviews. To suggest that such reviews do not touch upon subjective components of job assessments is clearly unrealistic. The model of subjective performance evaluation used in this paper and Fuchs (2007) should therefore be viewed as a simplification that can be improved further in future works.

1.5 Final remarks: Fuchs' mechanism vs. ours

Why is the money burning scheme $(0, 0, ..., 0, Z)$ proposed by Fuchs (2007) ideal in minimizing expected money burning but the same mechanism performs so poorly for the cost minimization objective? To understand the first part, let us start with an arbitrary money burning scheme: burn money z_0 upon the worst signal profile $\{\sigma_L \sigma_L \cdots \sigma_L\}$, burn z_j when there is only one high signal at the jth period $\{\sigma_L \sigma_L \cdots \sigma_H \cdots \sigma_L\}$, and for all other profiles burn zero money. For this scheme, assuming the agent exerts efforts in every period the expected money burning is given by $P_0 z_0 + P_j z_j$, where P_0 and P_j are the probabilities corresponding to the above two specific signal profiles.

For the above scheme *reuse of punishment* is not applicable, so we need to consider the jth period incentive besides how to deter $1st$ period deviation.[7] The marginal cost of shirking in the jth period is the difference between money burning upon $\{\sigma_L \sigma_L \cdots \sigma_L\}$ and $\{\sigma_L \sigma_L \cdots \sigma_H \cdots \sigma_L\}$, i.e. $z_0 - z_j$, times the increased probability of getting low signal in jth period. By lowering the amount z_j down to 0, and increasing z_0 by a small

[7] The reusable punishment idea was originally introduced in the repeated games literature by Abreu, Milgrom and Pearce (1991), which was used by Fuchs (2007) for his optimal mechanism construction. For reusable punishment, the principal only needs to ensure that the agent will not deviate to shirking in the first round which, in turn, guarantees that the agent will not deviate to shirking for any number of T rounds.

$\Delta > 0$, the marginal cost of shirking is pushed up, so that the agent will be more reluctant to shirk at the jth period, while keeping the other periods' incentives unchanged.[8] Now the expected money burning becomes $P_0(z_0+\Delta)+P_j\cdot 0 = P_0(z_0+\Delta)$, which is smaller than $P_0z_0+P_jz_j$ given Δ is small and the probability P_0 is also small.[9] Thus, modifying the incentives back towards Fuchs' mechanism with the reusability feature lowers expected money burning while implementing full efforts.

Also it is straightforward to see why Fuchs' mechanism fails for the cost minimization objective as already explained in the Introduction. Basically, instead of the very lop-sided punishment scheme of Fuchs, if money burning were spread out with less variance although in an increasing order according to the number of low signals, agent's effort incentives can be preserved and at the same time the maximum level of money burning can be brought down.[10]

[8] By manipulating the value of Δ, one is able to maintain the incentives for the first period, which is sufficient to support the equilibrium. This can be achieved analytically, and we skip the steps to keep the discussion short.

[9] Recall, P_0 is the probability of the worst signal profile $\{\sigma_L\sigma_L\cdots\sigma_L\}$ given full efforts, which is the lowest among all possible signal profiles so long as $p_1 > 1/2$.

[10] This will increase expected money burning relative to Fuchs' mechanism.

CHAPTER 2

Extreme vs. Moderate Wage Compression or Pay for Performance: Subjective Evaluation with Money Burning

2.1 Introduction

Most assessments by our superiors involve subjectivity and discretion. In fact when objective measures of performance are hard to obtain or not immediately available, employers must rely on subjective opinions or impressions of their subordinates' work to decide on the rewards: some assessment is better than no assessment and, as Baker et al. (1994) have argued, some element of subjectivity is better even when assessment can be made entirely objective. We ask how sensitive the rewards should be to performance when only subjective evaluation is possible.

We consider a principal-agent setting with agent moral hazard and subjective performance evaluation (*spe*). As is well known, under *spe* the principal has to ensure that he does not understate the agent's good performance, so he must be prepared to burn money. We will argue that, under appropriate assumptions, the optimal money burning scheme is either of the *pay for performance* type with the reward decreasing as performance drops (e.g., Holmstrom, 1979; Harris and Raviv, 1979), or one of *moderate wage compression* similar to Levin's (2003) *termination contract*.[1] The more *extreme wage compression*, where the agent

[1] Moderate wage compression typically involves, respectively, full and zero money

is penalized through money burning only when the performance is at its worst but otherwise receives a uniform reward, is more due to either agent's utility becoming unboundedly low (i.e., large negative) at very low (almost zero) consumption as shown in MacLeod (2003), or a principal maximizing social efficiency as in Fuchs (2007). Our twin results alluded to above, Propositions 2 and 6, open up new optimal contracting possibilities in the same environments considered by MacLeod (2003) and Fuchs (2007), and the results shift the balance, roughly, towards Levin's style of contracting – wage compression around a non-extreme threshold performance. Given perhaps the greater prevalence of this threshold based contracting in real life, this shift in results should improve our understanding of the wage compression hypothesis in principal-agent environments. Table 2.1 is a summary of various results. To place our paper properly in context, below we first review the related literature.

Table 2.1: ***spe* models: wage compression & pay for performance**[a]

	Profit max./cost min.	**Social efficiency/joint surplus max.**
Risk neutrality	Pay for performance or moderate wage compression: **this paper** (two-period game)[c]	Moderate wage compression: Levin (infinite repeated games) Extreme wage compression: Fuchs[b] (both finite & infinite repeated games)
Risk aversion	Extreme wage compression: MacLeod (one-shot game)	–x–x–x–x–x–x–
Risk aversion or Risk neutrality	Moderate wage compression: MacLeod 1-shot game minus $u(0) = -\infty$ **(this paper)**	–x–x–x–x–x–x–

[a] Wage compression: extreme = no money burning except for worst performance; moderate = money burning below a cutoff performance (above the worst).
[b] Our interest is in the finite repeated version.
[c] Chan and Zheng (2011) show pay for performance assuming 'no limited liability' of the agent that converts principal's obj. from profit max. to social efficiency.

burning below and above a threshold performance, and occasionally partial money burning at the threshold performance.

MacLeod (2003) studies a static principal-agent problem in which a risk-averse agent exerts a continuum of efforts in a project that yields a binary outcome, success or failure, not directly observable to anybody and contingent on outcome the effort translates into one of a finite number of performance signals (hinting at the project's likelihood of success) that is observed privately by the principal.[2] To provide effort incentives rewards must vary with performance but risk aversion should also limit the variability in rewards. What MacLeod finds, however, is quite striking: to maximize profits the principal ought to penalize the agent and burn money only when the performance is the worst possible, and for all other performance the rewards should be equalized that we refer as extreme wage compression (Proposition 6, MacLeod, 2003). While some amount of wage compression is natural, not penalizing at all for close-to-worst performance calls into question the power of incentives as one understands it from standard contract theory. We will see that such concentrated punishment has, surprisingly, nothing to do with the agent's risk aversion. Instead, an assumption of unbounded utility at zero consumption, along with a natural ordering on the informativeness of performance signals (monotone likelihood ratio condition), makes the specific flat reward structure optimal. Risk aversion should favor shifting some punishment towards better-than-the-worst but worse-than-the-best performance signals. But this economic reasoning doesn't seem to have any pivotal role. On the other hand, if the unbounded utility assumption is dropped, irrespective of whether the agent is risk averse or risk neutral, the optimal rewards structure will move away from MacLeod-postulated extreme compression.

Levin (2003) shows that the optimal contract in an infinite repeated principal-agent setting with moral hazard and *spe* involves a *one-step*

[2] The author also considers the case where the agent might observe a signal that is correlated with the principal's information.

termination contract:[3] a base wage w with contract termination if privately observed performance level, y_t, falls below a threshold level $\hat{y}$, or continuation with an additional bonus b if $y_t \geq \hat{y}$ (Proposition 7);[4] this pattern we refer as moderate wage compression to distinguish it from extreme wage compression. The agent in Levin's analysis is risk neutral. The repeated relationship, through continuation values, helps to endogenize money burning triggered by costly disputes and termination of the relationship.

Fuchs (2007), like Levin (2003), studies a repeated principal-agent game where in each round the agent either exerts one unit of effort or shirks. Assuming that the principal wants to minimize (expected) money burning the author shows that when the repeated game involves a finite number of T rounds, in each of which the (risk-neutral) agent should be induced to exert effort, the principal should burn money only when the privately observed evidence of agent performance in all T rounds are low. This, again, is a form of extreme wage compression in the mould of MacLeod (2003): burn money a lot but very infrequently or otherwise don't burn money at all.[5,6]

Our principal-agent models borrow some features of the above three studies and departs in others. The first of two models to be studied is a simplified version of Fuchs (2007), but permits the more noisy performance evaluation of MacLeod (2003).[7] A principal hires an agent to work

[3] Levin defines a self-enforcing incentive program to be optimal if it maximizes per-period expected joint surplus of the principal and the agent. An incentive program (or contract) specifying agent compensation for all possible histories is self-enforcing if it induces Nash equilibrium play of the infinite repeated game following each history.

[4] Levin addresses an even more general problem with the additional issue of adverse selection. Our comparison is with the simpler version of his analysis.

[5] Fuchs also considers the infinite repeated version.

[6] Prendergast (1993) and Prendergast and Topel (1996) also analyze subjective evaluation – an agent reports information relevant for the principal's decision which is evaluated against principal's own information – and sometimes the optimal contract breaks down agent performance into two categories, acceptable and unacceptable, thus exhibiting compression in (agent) rating.

[7] Fuchs' (2007) model, in turn, shares some features of Levin (2003).

over two periods in each of which the agent either exerts effort or shirks. The principal wants to implement full efforts over two rounds at minimal cost by promising rewards contingent on his subjective assessment of the agent's performance in each round; either he directly observes a nonverifiable output performance or a signal of performance. By restricting to two periods we keep the analysis tractable but it also reflects the fact that most employment relations are of finite length. This is the first point of departure from Levin (2003) and to an extent Fuchs (2007). By not allowing infinitely long relationship our model will not be able to endogenize money burning in the way Levin (2003) does. Our principal will thus use money burning directly as an incentive instrument.[8] Organizations rarely deal with only a single agent, thus money burning to discipline one agent can always be passed onto another agent or some other department within the organization, an interpretation that is both realistic and similar in spirit to MacLeod's (2003) interpretation that the burnt money is given to a "third-party" (see p. 222).[9] In a second formulation, we use MacLeod's (2003) static game but drop the assumption of agent ruin near zero consumption by assuming utility bounded from below and broaden the applicable preferences to allow for utility of money to be linear as a second possibility (i.e., $u''(\cdot) \leq 0$). Third, different from Fuchs (2007) and Levin (2003) but more like MacLeod (2003), our principal minimizes his reward costs (or maximizes profit) rather than maximizing social efficiency or joint surplus.

We will see that the above modelling differences will combine to yield, often, the *pay for performance* incentives (Proposition 5). This conforms to our casual understanding that if the year-end meeting between an employer and an employee comes to conclude that the agent has not

[8] This was also the case in MacLeod (2003), in the finite repeated game model of Fuchs (2007), and Chan and Zheng (2011).

[9] The use of bonus pools to incentivize a group of employees is a standard practice (Rajan and Reichelstein, 2006; 2009).

performed well over a certain period by whatever subjective assessment conducted by the employer, the compensation is likely to reflect in adjusted salaries and/or bonuses in relation to the degree of underperformance. This is especially so if the organization has a fixed salary or bonus pool that must be distributed in an equitable manner across its employees. Although we are not explicitly modelling the determination of reward of an agent within a group, the incentives of an agent can be viewed informally in an employment setting involving other employees. Empirically, 'performance pay' under subjective assessment has been known to perform well (Kahn and Sherer, 1990).[10,11] Thus, variable pay and bonuses can be optimal outside the earlier hypothesis of employer bias or arbitrary discretion (Prendergast, 1993, 1999; Prendergast and Topel, 1996).

The simple modification of MacLeod's (2003) contracting game by dropping the Inada condition leads to a softening of his extreme wage compression hypothesis (Proposition 6). This result brings performance-pay back into play and makes wage compression *moderate* by extending full money burning beyond the worst performance scenario. As noted earlier, this wage compression, which is perhaps more realistic, is similar

[10] Murphy and Oyer (2003) find, while evaluating the costs and benefits of subjective performance evaluation relative to objective measures, that discretion is more important in determining executive bonuses at larger and privately held firms.

[11] Governments in the United Kingdom and at various state levels in the USA are increasingly relying on performance-related pay for teachers, where authorities assess teacher effectiveness from student grades but also based on other criteria that can introduce subjectivity. See the UK government press release on 29 April, 2013: "New advice to help schools set performance-related pay" (https://www.gov.uk/government/news/new-advice-to-help-schools-set-performance-related-pay). It states: "The advice published today highlights factors schools could consider when assessing teachers performance. This includes a teacher's impact on pupil progress, impact on wider outcomes for pupils, contribution to improvements in other areas (e.g. pupils' behaviour or lesson planning), professional and career development, wider contribution to the work of the school, for instance their involvement in school business outside the classroom. Schools could consider evidence from a range of sources, including self-assessment, lesson observations, and the views of other teachers and of parents and pupils." For the USA, see http://www.latimes.com/local/teachers-investigation/#axzz2ut1ScDxw; http://files.eric.ed.gov/fulltext/ED535859.pdf; and the works of Neal (2011), and Neal and Barlevy (2012).

to Levin's (2003) result but is obtained in a static game with money burning used as an instrument.[12] Explaining wage compression in a static game should be a useful exercise given that most relations are of finite duration.[13] Thus, Proposition 6 should be seen as complementary to MacLeod (2003), further expanding the reach of his model.

We do a robustness check of the performance-pay hypothesis when the agent observes a signal correlated with the principal's information (Proposition 7). For this test we follow Chan and Zheng (2011), who study principal-agent dynamic moral hazard and contracting under *spe* and show a similar performance-pay result under the restrictive assumption that the agent is not subjected to 'limited liability'. Their analysis is equivalent to a principal maximizing social efficiency (as opposed to our cost-minimizing principal) and they show that the principal should reward an improving performance trajectory more than a declining trajectory. Our analysis yields performance-pay without necessarily the bias due to specific upward or downward trajectory identified by Chan and Zheng (Proposition 8). Further, to suggest that our performance-pay result is not an anomaly due to the specific two-period formulation, through numerical simulation we demonstrate how performance pay can dominate extreme wage compression in a three-period game; see Fig. 2.1.

The rest of the paper is organized as follows. In the next section we present the model. Our main analysis and the results are contained in sections 3-5, with conclusions appearing in section 6. Proofs are in-

[12] Other explanations of wage compression are in the more traditional setting of firms determining relative pay of employees; see, for example, Lazear (1989), and Fang and Moscarini (2005). Lazear argues that equal pay reduces non-cooperation and sabotage within the organization, whereas Fang and Moscarini exploit the theme of workers' low morale (or confidence) following revelation of their true ability as most workers are overconfident and so through wage non-differentiation employers can perpetuate workers' misperception and maintain a positive attitude to work.

[13] Much of the insight for a finite repeated agency model can be derived from our modified static game analysis: the common thread between Propositions 5 and 6 should become clearer.

cluded in Appendix.

2.2 The model

A risk-neutral principal involves a risk-neutral agent in a two-period repeated efforts game. In each period the agent can either exert one unit of effort or shirk, $e_t \in \{0, 1\}$, which is not observable to the principal, with effort costing the agent $c > 0$. There is no discounting by the principal or the agent. The output in each period can be either high or low, $y_t \in \{y_L, y_H\}$, depending on the effort in that period:

$$\Pr[y_t = y_H \mid e_t = 0] = \beta_0, \quad \Pr[y_t = y_H \mid e_t = 1] = \beta_1.$$

Instead of directly observing output, the principal may observe only some signal of agent performance, $\sigma_t \in \{\sigma_H, \sigma_L\}$, which is private and cannot be disclosed verifiably. The probability of σ_t given the output is high is $\gamma^H_{\sigma_t}$, and given the output is low it is $\gamma^L_{\sigma_t}$. Given the binary signals in each state, we have $\gamma^H_{\sigma_H} + \gamma^H_{\sigma_L} = \gamma^L_{\sigma_H} + \gamma^L_{\sigma_L} = 1$. Further, the following *monotone likelihood ratio* condition (Milgrom, 1981) will be assumed:

$$\frac{\gamma^H_{\sigma_H}}{\gamma^L_{\sigma_H}} > \frac{\gamma^H_{\sigma_L}}{\gamma^L_{\sigma_L}}.$$

Therefore, during a particular round the probability of observing any performance signal depends only on the effort in that round as follows:

$$p_0 \equiv \Pr[\sigma_t = \sigma_H \mid e_t = 0] = \beta_0 \gamma^H_{\sigma_H} + (1 - \beta_0)\gamma^L_{\sigma_H},$$
$$p_1 \equiv \Pr[\sigma_t = \sigma_H \mid e_t = 1] = \beta_1 \gamma^H_{\sigma_H} + (1 - \beta_1)\gamma^L_{\sigma_H},$$

with $0 < p_0 < p_1 < 1$.

Suppose the principal's expected wage payment for two periods is

$\mathbb{E}(W)$, then his profit is given by:

$$\Pi = \sum_t \mathbb{E}(y_t \mid e_t) - \mathbb{E}(W).$$

Let $\mathbf{e} = (e_1, e_2)$ be the agent's efforts over two rounds. To solve the principal's problem, the first step is to minimize his cost $\mathbb{E}(W)$ of inducing any effort profile $\mathbf{e}$; then the second step is to determine the optimal $\mathbf{e}^*$ to maximize the profit Π. To simplify our analysis, we assume that the incremental expected output in any period following change from $e = 0$ to $e = 1$ is large enough so that the principal wants to uniquely implement $\mathbf{e}^* = (1, 1)$ at minimal wage costs.

If the principal observes the output, as assumed by Fuchs (2007), then the reward scheme and the principal's expected wage cost should be contingent on the output profile $\mathbf{y} = (y_1, y_2)$; if the principal observes only the performance signal, as assumed by MacLeod (2003), then his cost should be a function of the signal profile $\mathrm{s} = (\sigma_1 \sigma_2)$. Throughout our analysis we use the latter formulation but the first interpretation is also possible.

At the end of two periods the principal will report the performance signals s. Let $w^P(\mathrm{s})$ be the amount of money pre-committed to be spent by the principal for any reported profile of performance. Due to the subjective performance evaluation, the principal has to pay a fixed amount regardless of his private observation of signals in order for him not to misreport the agent's performance. However, to incentivize the agent to exert effort, the reward should be contingent on the reported performance. Thus, the agent should not always receive all the money paid by the principal; under some circumstances the budget balance may break down, which formally amounts to *money burning*.[14] Denote the principal's (fixed) budget $w^P(\mathrm{s}) = W$, the agent's reward as $w^A(\mathrm{s})$, and

[14] MacLeod (2003) proposed the idea of money burning under *spe*.

contingent money burning as $z(\mathrm{s})$. Let the set of all signal profiles be $\Sigma = \{\mathrm{s}\} = \{(\sigma_1\sigma_2)\}$, and bold symbols $\mathbf{w^A}$ and $\mathbf{z}$ be vectors of rewards and money burning corresponding to different signal profiles s. Formally, the payment scheme is defined as:

$$\begin{aligned} &\left(W, \mathbf{w^A}(\Sigma), \mathbf{z}(\Sigma)\right), \qquad \text{where} \\ &W = w^A(\mathbf{s}) + z(\mathbf{s}) \qquad \forall \mathbf{s} \in \Sigma. \end{aligned} \tag{2.1}$$

We may simply write $(W, \mathbf{z}(\Sigma))$ to refer to the incentive mechanism.

Given the incentives, the agent's expected utilty (or payoff) of exerting effort profile e can be written as follows:

$$\begin{aligned} V(\mathbf{e}) &= \mathbb{E}\left(w^A(\mathbf{s}) \mid \mathbf{e}\right) - c\sum_t e_t \\ &= W - \sum_{\mathbf{s}\in\Sigma} z(\mathbf{s})\ \Pr(\mathbf{s} \mid \mathbf{e}) - c\sum_t e_t. \end{aligned} \tag{2.2}$$

Then the principal solves the following problem:

$$\min_{W, \mathbf{z}(\Sigma)} \quad W \tag{2.3}$$

$$\text{s.t.} \quad \text{(Incentive Compatibility)} \quad \mathbf{e}^* \in \arg\max V(\mathbf{e}), \tag{2.4}$$

$$\text{(Limited Liability)} \quad W - z(\mathbf{s}) \geq 0 \quad \forall \mathbf{s} \in \Sigma, \tag{2.5}$$

$$\mathbf{z}(\Sigma) \geq \mathbf{0}. \tag{2.6}$$

Incentive compatibility (2.4) and limited liability (2.5) will guarantee agent's participation constraint $V(\mathrm{e}^*) \geq 0$.

We assume no interim feedback or payment/money burning. The time line of the game is as follows:

1. *At time zero, a contract* $(W, \mathbf{z}(\Sigma))$ *is signed between parties;*

2. *At each period* $t = 1, 2$*, the agent decides whether to exert effort*

on the principal's project or shirk and the principal receives a performance signal at the end of the period;

3. *At the end of two periods, the principal reports the performance signal profile* $\mathbf{s}$ *and makes the payment* $w^A(\mathbf{s})$.

The model presented above is a two-period variant of Fuchs' (2007) finite period model with a more noisy subjective assessment as in MacLeod (2003). The more important difference, however, is in the specification of the principal's objective – he minimizes effort implementation costs rather than maximizing social efficiency. This change is a natural one given our interest in an incentive mechanism for a profit-seeking principal. To implement efforts the principal should like to minimize his own costs. As we will see, the different objectives may lead to a sharp difference in the optimal contracts.

2.3 Pay for performance or moderate wage compression vs. extreme wage compression

In this section, we solve for the optimal money burning contract for a profit-motivated principal and show that the optimal contract often exhibits the pay-for-performance principle.

As analyzed in Chapter 1, solving the principal's problem (2.3) subject to (2.4)-(2.6) yields the following characterization of the optimal contract:

PROPOSITION 5 (**Optimal contract: complete characterization**). *Under subjective evaluation a profit-seeking principal often relies on the* pay for performance *principle or* moderate wage compression*, although in some situations* extreme wage compression *is still a possibility. More specifically, the principal's optimal contract is as follows:*

(I) *[Moderate compression] If $p_1 > \frac{1}{2}$ and $p_1 + p_0 > 1$,*

$$W = \tfrac{2c}{(p_1+p_0)(p_1-p_0)}, \;\; z^{HH} = 0, \;\; z^{LH} = z^{HL} = z^{LL} = \tfrac{2c}{(p_1+p_0)(p_1-p_0)}.$$

(II) *(a) [Pay for performance/moderate compression] If $p_1 > \frac{1}{2}$ and $p_1 + p_0 = 1$,*

$$W = \tfrac{2c}{p_1-p_0}, \;\; z^{HH} = 0, \;\; z^{LL} = \tfrac{2c}{p_1-p_0},$$

$z^{LH}, z^{HL} \in [\frac{c}{p_1-p_0}, \frac{2c}{p_1-p_0}]$, and $p_1 z^{HL} \geq p_0 z^{LH} + c$, $p_0 z^{HL} \geq p_1 z^{LH} - 2c$.

(b) [Pay for performance] If $p_1 > \frac{1}{2}$ and $p_1 + p_0 < 1$,

$$W = \tfrac{2c}{p_1-p_0}, \;\; z^{HH} = 0, \;\; z^{LH} = z^{HL} = \tfrac{c}{p_1-p_0}, \;\; z^{LL} = \tfrac{2c}{p_1-p_0}.$$

(c) [Pay for performance/extreme compression] If $p_1 = \frac{1}{2}$,

$$W = \tfrac{2c}{p_1-p_0}, \;\; z^{HH} = 0, \;\; z^{LH} = z^{HL} \in [0, \tfrac{c}{p_1-p_0}], \;\; z^{LL} = \tfrac{2c}{p_1-p_0}.$$

(III) *[Extreme compression] If $p_1 < \frac{1}{2}$,*

$$W = \tfrac{c}{(1-p_1)(p_1-p_0)}, \;\; z^{HH} = z^{LH} = z^{HL} = 0, \;\; z^{LL} = \tfrac{c}{(1-p_1)(p_1-p_0)}.$$

Note that given the partition of the signal generating technology, $[0,1] \times [0,1]$, the conditions stated above are actually a complete, *if and only if*, characterization of the optimal contract.

Next, let us explain why extreme wage compression need no longer be an optimal choice for the principal. In Fuchs (2007), like ours, the agent had to be induced to exert efforts but the principal's main concern was to do so at minimal expected money burning. Our principal, on the other hand, is interested in minimizing his own reward cost that equals the *maximum* of the burnt money. Since minimizing expected money burning is not necessarily the dual of the principal's profit maximization problem, solution to Fuchs' implementation program does not necessarily minimize principal's costs.

Let us now return to the economic intuitions for different wage compression schemes for different constellations of parameters (or signal generating technology).

When is *moderate wage compression* optimal? This is the first case in Proposition 5, and can be described, paradoxically, as the rule of: *"one strike and you are out"*.[15] Taking the heavy punishment scheme as **given**, let us first see when this scheme is likely to uniquely implement $\mathbf{e} = (1, 1)$. Intuitively, such punishment must imply that the probability of wrongfully penalizing the diligent agent (for suspicion of shirking) is less, and perhaps considerably so, than the probability of mistakenly letting the shirking agent get away free. To see when this might be true, consider pr(at least one low signal$|\mathbf{e} = (1, 1)) = (1-p_1)^2 + 2p_1(1-p_1)$; pr(at least one low signal$|\mathbf{e} = (0, 0)) = (1-p_0)^2 + 2p_0(1-p_0)$; and pr(at least one low signal$|\mathbf{e} = (1, 0)) = (1 - p_1)p_0 + (1 - p_0)p_1 + (1 - p_1)(1 - p_0)$. Now suppose (i) pr(at least one low signal$|\mathbf{e} = (1, 1))$ is very low, (ii) pr(at least one low signal$|\mathbf{e} = (0, 0))$ is high, and (iii) pr(at least one low signal$|\mathbf{e} = (1, 0))$ is such that the agent would rather switch from $\mathbf{e} = (1, 0)$ to $\mathbf{e} = (1, 1)$ than to $\mathbf{e} = (0, 0)$. The first two conditions will ensure that the agent should choose $\mathbf{e} = (1, 1)$ over $\mathbf{e} = (0, 0)$, which can happen only if $(1 - p_1)^2 + 2p_1(1 - p_1) < (1 - p_0)^2 + 2p_0(1 - p_0)$, which reduces to: $p_1 + p_0 > 1$. Coming to requirement (iii), we must have pr(at least one low signal$|\mathbf{e} = (1, 1))$ "sufficiently" less than pr(at least one low signal$|\mathbf{e} = (1, 0))$, which, in turn, must be "sufficiently"

[15] Moderate wage compression can be of a less extreme nature, in principle, when contract spans over more than two periods. With a two-period contract, overall signals can contain only one or two low signals and thus there is not much room for varied types of moderate wage compression – either the contract is of moderate wage compression (uniform money burning with one low signal or more) or agent compensation is strictly improving in performance. In section 4, with any number of signals, $n > 2$, moderate wage compression compensation can be a bit more forgiving.

less than pr(at least one low signal$|\mathbf{e} = (0,0))$. These last two will hold, it is easy to verify, only if $p_1 - p_0 > 0$ is sufficiently high, which, combined with the requirement that $p_1 + p_0 > 1$, implies $p_1 > \frac{1}{2}$. Thus, moderate wage compression implementing $\mathbf{e} = (1,1)$ uniquely, implies $p_1 + p_0 > 1$ and $p_1 > \frac{1}{2}$.

Now to understand why with $p_1 > \frac{1}{2}$ and $p_1 + p_0 > 1$ the principal should adopt flat and heavy punishment, note that either p_1 must be quite high especially when p_0 is low, or both p_1 and p_0 are fairly high. If p_1 is close to 1 say, exerting efforts will generate high signals almost surely, so low signal in any round is a clear indication of shirking in that round rather than bad luck; if both p_1 and p_0 are high, shirking also has a good chance of generating a high performance signal. In either of these two signal generating scenarios, there is little to differentiate between one low signal and two low signals, such are the smallness of their respective likelihoods. This implies it might not be in the principal's interest to make a nuanced differentiation between one and two low signals, and hence he should penalize the agent uniformly. And the maximal punishment follows because with two low signals the principal must come down on the agent in the strongest possible manner and not leave the agent with any positive surplus ex post.

The intuition for *performance pay* can be understood as follows. When $p_1 > \frac{1}{2}$ and $p_1 + p_0 \leq 1$, it implies that the difference in the probabilities of generating a high signal when the agent exerts effort vs. when he shirks is going to be non-trivial ($p_1 > \frac{1}{2}$ but $p_0 < \frac{1}{2}$), thus the choice of effort or shirking is very likely to be reflected in the signal generated: signals, although imperfect, are informative and hence the scope for nuanced rewards/punishment; all money is burnt if both periods see the low signals, whereas partial money is burnt upon a combination of one low and one high signal.

Finally, while the *extreme wage compression* remains a possibility, the environment might be considered less natural. When $p_1 \leq \frac{1}{2}$, the high signal is less likely than the low signal with the agent exerting effort. This is a case of weak informativeness of high signal. While high signal does not suggest a strong evidence of agent's effort, low signal, on the other hand, would imply a high chance that the agent did *not* exert effort: $p_0 < \frac{1}{2}$. That is, rather than the high signal, its *absence* is more indicative of lack of effort. With $p_0 < p_1$, the principal can rely on the signals' informativeness (as monotone likelihood ratio property [Milgrom, 1981] will be satisfied) to determine money burning. The principal chooses to burn money only when signals in both periods are low; when only one signal is low, it could be that the agent did exert effort in that round yet he was unlucky (recall $p_1 \leq \frac{1}{2}$) and the principal does not want to wrongfully penalize the agent.

To summarize, the first two cases burn money whenever there is at least one low signal. This represents what we may call, broadly, the *pay for performance* principle, and is similar to standard contracts with verifiable performance (Holmstrom, 1979; Harris and Raviv, 1979).[16] The second case is more discriminating with the agent's penalty (or money burning) strictly increasing in the number of low signals. And the last case is same as the *extreme wage compression* result of MacLeod (2003) and Fuchs (2007).[17]

In view of Proposition 5, it is not unreasonable to suggest that the extreme wage compression result in Fuchs (2007) was largely driven by the principal's social efficiency maximization hypothesis. The two-period

[16] There is some parallel with the familiar law-and-econ doctrine of "penalty fitting the crime" (Andreoni, 1991; Shavell, 1991). One should recognize though that money burning is never an issue for the law-and-econ doctrine. There the main concern is fairness but equally important is the implication for deterrence of more serious crimes.

[17] When $T = 2$ in Fuchs (2007), his expected money burning minimization contract is given by: $W = \frac{c}{(1-p_1)(p_1-p_0)}$, $z^{HH} = z^{LH} = z^{HL} = 0$, $z^{LL} = \frac{c}{(1-p_1)(p_1-p_0)}$.

model is a special case of Fuchs' T-period model, and the only change in our analysis is in the principal's objective from social efficiency to profit maximization (equivalently, cost minimization).

■ Pay for performance in a three-period model: an illustration. Instead of the extreme wage-compression scheme, if money burning were spread out with less variance although in an increasing order according to the number of low signals, agent's effort incentives can be preserved and at the same time the maximum level of money burning, and thus principal's cost, can be brought down. Below we provide a numerical illustration of a three-period contract.

Let $Z \equiv (z(1), z(2), z(3))$ be a money burning scheme, where $z(i)$ refers to the amount of money burning upon $\#i$ observation(s) of low signal(s). Then it is sufficient to consider the agent's decision of exerting effort for how many periods out of three, regardless of the order. The incentive compatibility conditions for the agent $V(1,1,1) \geq V(1,1,0)$, $V(1,1,1) \geq V(1,0,0)$ and $V(1,1,1) \geq V(0,0,0)$ where $V(\cdot,\cdot,\cdot)$ is defined similar to $V(\cdot,\cdot)$, can be written in terms of expected money burning and effort costs, given the fixed wage paid by the principal. Explicitly,

$$\begin{aligned}
&3p_1^2(1-p_1)z(1) + 3p_1(1-p_1)^2 z(2) + (1-p_1)^3 z(3) + 3c \\
&\quad \leq [p_1^2(1-p_0) + 2(1-p_1)p_1p_0]z(1) + [p_0(1-p_1)^2 + 2p_1(1-p_1)(1-p_0)]z(2) + (1-p_1)^2(1-p_0)z(3) + 2c; \\
&3p_1^2(1-p_1)z(1) + 3p_1(1-p_1)^2 z(2) + (1-p_1)^3 z(3) + 3c \\
&\quad \leq [p_0^2(1-p_1) + 2(1-p_0)p_0p_1]z(1) + [p_1(1-p_0)^2 + 2p_0(1-p_0)(1-p_1)]z(2) + (1-p_0)^2(1-p_1)z(3) + c; \\
&3p_1^2(1-p_1)z(1) + 3p_1(1-p_1)^2 z(2) + (1-p_1)^3 z(3) + 3c \\
&\quad \leq 3p_0^2(1-p_0)z(1) + 3p_0(1-p_0)^2 z(2) + (1-p_0)^3 z(3).
\end{aligned}$$

Let Z_1 be a wage-compression scheme so that $Z_1 = (0, 0, z_1(3))$. Consider Z_2 to be the money burning scheme whenever at least two low signals are observed, i.e. $Z_2 = (0, z_2(2), z_2(3))$, and Z_3 be the money burning scheme whenever at least one low signal is observed so that $Z_3 = (z_3(1), z_3(2), z_3(3))$. Set the following parameter values: $p_0 = 0.25$, $p_1 = 0.625$ and $c = 1$. Then using *Mathematica* the following sample solutions are generated and plotted in Fig. 2.1:

	$z(1)$	$z(2)$	$z(3)$	$W = \max\{z(i)\}$
Z_1	0	0	18.96	18.96
Z_2	0	2.56	13.00	13.00
Z_3	3.19	4.29	10.72	10.72

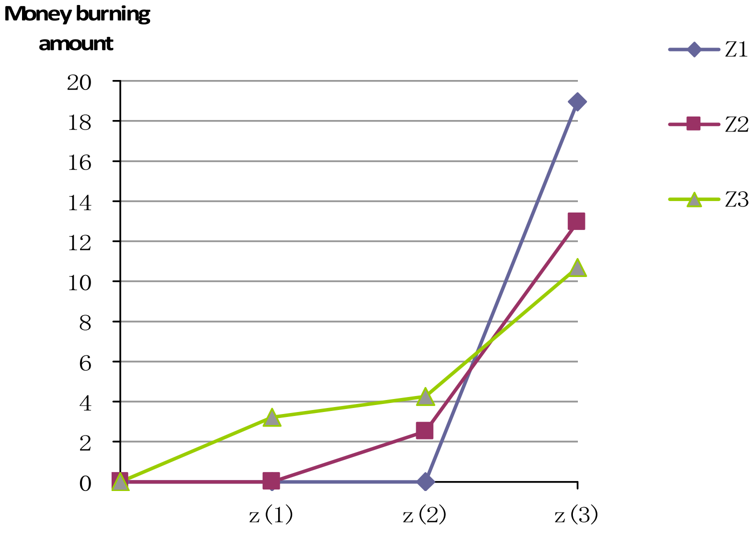

Figure 2.1: Different money burning schemes

The above example shows that our pay-for-performance hypothesis is applicable beyond the two-period model studied. So long as the agency relationship is of finite length, similar incentives linking pay to performance (via the signals) should work because money burning takes care of the principal's incentives to misreport.

2.4 Beyond MacLeod (2003): More general risk preferences, $u'(0) < \infty$, and moderate compression

In this section, we revisit MacLeod (2003) to point out the critical role of the assumption that the agent is ruined when consumption is very low for his extreme wage compression result. We are going to argue that if instead the agent's utility is bounded below so that the Inada condition is violated, MacLeod's mechanism will be dominated by more moderate wage-compression incentives.

MacLeod (2003) considers a static principal-agent game in which the agent privately chooses an effort $\lambda \in [0, 1)$ into a project leading to its success (state H) with probability λ and failure (state L) with probability $1 - \lambda$. The principal observes only a non-verifiable signal $t \in \mathcal{T}$ of the agent's performance, with $|\mathcal{T}| = n > 2$. The probability of signal t realization given effort λ is $\gamma_t(\lambda) = \lambda\gamma_t^H + (1 - \lambda)\gamma_t^L$, where γ_t^H (or γ_t^L) is the probability of signal t given that the project is a success (or failure).

At the time of signing the contract, the principal commits to spend $w_t = \bar{w}$ as potential reward part of which, b, will be burnt if the performance signal t is not satisfactory. The agent receives $c_t = \bar{w} - b$.

MacLeod further makes the following assumptions:

Assumption 1-MacLeod: The Bernoulli utility function of the agent satisfies $U(c, \lambda) = u(c) - V(\lambda)$, where $c > 0$, $\lambda \in [0, 1)$ and $u' > \epsilon > 0$, $u'' < 0$, $\lim_{c \to 0} u(c) = -\infty$, $V' > 0$, $V'' > 0$ and $\lim_{\lambda \to 1} V(\lambda) = \infty$. That is, the agent is risk averse (utility is concave in consumption) and effort cost is convex.

Assumption 2-MacLeod (Generic monotone likelihood ratio condition): $\gamma_{t+1}^H/\gamma_{t+1}^L > \gamma_t^H/\gamma_t^L > 0$ for $t = 1, \ldots, n - 1$, and for no t is it the case that

$\gamma_t^H/\gamma_t^L = 1$.

MacLeod then shows the following result (the descriptive title added by us).

Proposition 6-MacLeod (2003) (**Penalty as a last resort & extreme wage compression**). *Suppose Assumptions 1 and 2 in MacLeod (2003) hold, and there is no correlation in the principal's and the agent's beliefs. Then the optimal contract implementing any effort $\lambda \in (0,1)$ based on subjective evaluations entails wage payments that do not depend upon the principal's evaluation:*

$$w_t = \bar{w} \quad \forall t \in \mathcal{T},$$

while the agent receives:

$$c_t = \begin{cases} \bar{w}, & \textit{if} \quad t > 1 \\ \bar{w} - b, & \textit{if} \quad t = 1 \end{cases} \tag{2.7}$$

where $0 < b < \bar{w}$ and $t = 1$ corresponds to the lowest performance level, i.e., the signal with the lowest likelihood ratio among all γ_t^H/γ_t^L.[18]

We now drop the assumption that $\lim_{c\to 0} u(c) = -\infty$ and replace it by $\lim_{c\to 0} u(c) > -\infty$ and $\lim_{c\to 0} u'(c) < \infty$, and further assume $u''(c) \leq 0$ (i.e., utility of money can be linear).[19] This opens up more room for performance related information to influence agent compensation.

Given that $\gamma_t^H/\gamma_t^L \neq 1$, the set of signals $\mathcal{T}$ can be partitioned as follows:

$$\mathcal{T} = \mathcal{T}^+ \bigcup \mathcal{T}^-,$$

[18] In MacLeod (2003), optimal $\bar{w}$ and b are determined solving the incentive and participation constraints, both of which are binding.

[19] Thus, the agent can be risk averse or risk neutral – a generalization of the risk preference of MacLeod's agent.

where $\mathcal{T}^+ = \{t : \gamma_t^H - \gamma_t^L > 0\}$ and $\mathcal{T}^- = \{t : \gamma_t^H - \gamma_t^L < 0\}$. Further, by monotone likelihood ratio condition, we have:

LEMMA 1.

(i) $\frac{\gamma_t^H - \gamma_t^L}{\gamma_t(\lambda)}$ *is strictly increasing in* t.

(ii) *If for some* $t = \hat{t}$, $\gamma_{\hat{t}}^H - \gamma_{\hat{t}}^L < 0$*, then for all* $t < \hat{t}$, $\gamma_t^H - \gamma_t^L < 0$.

Proof of (i) follows applying $MLRC$ to: $\frac{\gamma_t^H - \gamma_t^L}{\gamma_t(\lambda)} = \frac{\gamma_t^H - \gamma_t^L}{\lambda\gamma_t^H + (1-\lambda)\gamma_t^L} = \frac{1}{\lambda + \frac{\gamma_t^L}{\gamma_t^H - \gamma_t^L}} = 1/[\lambda + 1/(\frac{\gamma_t^H}{\gamma_t^L} - 1)]$. Then (ii) follows from (i).

From Lemma 1,

$$\mathcal{T}^- = \{1, 2, \ldots, K\} \text{ and } \mathcal{T}^+ = \{K+1, K+2, \ldots, n\},$$

where at $t = K$, $\gamma_K^H - \gamma_K^L < 0$ and at $t = K+1$, $\gamma_{K+1}^H - \gamma_{K+1}^L > 0$.

LEMMA 2. *The principal should burn money only upon signals such that* $\gamma_t^H < \gamma_t^L$*, i.e.*

$$\mathcal{T}^{\mathcal{MB}} = \{t : \textit{money burning} > 0\} \subseteq \mathcal{T}^-.$$

Proof. Suppose the contract implementing effort λ involves money burning at some signal $\check{t} \in \mathcal{T}^+$, implying $c_{\check{t}} < \bar{w}$, where $\bar{w}$ is the fixed budget for the principal. Suppose also that both the incentive compatibility and participation constraints are binding:

$$\text{[Incentive Constraint]} \quad \sum_{t=1}^{n} u(c_t)(\gamma_t^H - \gamma_t^L) - V'(\lambda) = 0,$$
$$\text{[Participation Constraint]} \quad \sum_{t=1}^{n} u(c_t)\gamma_t(\lambda) - V(\lambda) - \bar{u} = 0.$$

Then increasing $c_{\check{t}}$ by $\Delta = \bar{w} - c_{\check{t}} > 0$ will make both constrains relaxed since $\gamma_{\check{t}}^H - \gamma_{\check{t}}^L > 0$ and $\gamma_{\check{t}}(\lambda) > 0$. Now start lowering $\bar{w}$ and with it any c_t that was previously set at $c_t = \bar{w}$ until one of the two constraints, IC and

PC, binds, at which point λ is implemented with a new lower cost for the principal. This means the original contract with $c_{\check{t}} < \bar{w}$ could not have been optimal. **Q.E.D.**

The following result contrasts with the extreme wage compression result of MacLeod (2003). The optimal contract is one of *moderate wage compression*: wages are uniform in two segments separated by a threshold performance.

PROPOSITION 6 (**2-fold wage compression**). *Consider MacLeod's (2003) static* spe *contracting model with a more general $u(c)$, where $u''(c) \leq 0$ so that utility can even be linear in c. Further, different from MacLeod, assume that $\lim_{c\to 0} u'(c) < \infty$ (violation of the Inada condition) and $u(0) = 0$.*[20] *Then the optimal contract implementing any effort $\lambda \in [0, 1)$ can be characterized as follows:*

(i) If $\frac{V'(\lambda)}{V(\lambda)+\bar{u}} \leq \frac{\sum_{t=K+1}^{n}(\gamma_t^H-\gamma_t^L)}{\sum_{t=K+1}^{n}\gamma_t(\lambda)}$,

$$c_t = \begin{cases} 0 & \text{if } \ t < t', \\ \bar{w} - b & \text{if } \ t = t', \\ \bar{w} & \text{if } \ t > t', \end{cases}$$

where $0 < b \leq \bar{w}$ and a unique $t' \in \mathcal{T}^-$ binding the agent's incentive and participation constraints.

(ii) If $\frac{V'(\lambda)}{V(\lambda)+\bar{u}} > \frac{\sum_{t=K+1}^{n}(\gamma_t^H-\gamma_t^L)}{\sum_{t=K+1}^{n}\gamma_t(\lambda)}$,

$$c_t = \begin{cases} 0 & \text{if } \ t \in \mathcal{T}^-, \\ \bar{w} & \text{if } \ t \in \mathcal{T}^+, \end{cases}$$

[20] This is equivalent to dropping MacLeod's condition that $\lim_{c\to 0} u(c) = -\infty$ and $\lim_{c\to 0} u'(c) = \infty$. Also note that our assumption that $u(0) = 0$ can be replaced by $u(0) > -\infty$, without changing any of the results. The normalization simplifies the proof.

with the agent's incentive constraint binding and a slack in the participation constraint (so that the agent earns a rent).

(The precise values of $u(\bar{w})$ and $u(\bar{w}-b)$ are derived explicitly in the proof.)

Proof. To implement any effort $\lambda \in [0,1)$, the principal's problem can be stated as follows:

$$\begin{aligned}
&\min_{\bar{w},\{c_t\}} \quad \bar{w} \qquad\qquad \text{subject to:}\\
\text{[Incentive Constraint]}\quad &\sum_{t=1}^{n} u(c_t)(\gamma_t^H - \gamma_t^L) - V'(\lambda) \geq 0,\\
\text{[Participation Constraint]}\quad &\sum_{t=1}^{n} u(c_t)\gamma_t(\lambda) - V(\lambda) - \bar{u} \geq 0,\\
&\bar{w} - c_t \geq 0 \qquad \forall\, t,\\
\text{[Non-negativity Constraint]}\quad &c_t \geq 0 \qquad \forall\, t,
\end{aligned}$$

where $\bar{u}$ is the value of the agent's outside option.

Compared with the principal's problem in MacLeod (2003), we have the additional non-negativity constraints $c_t \geq 0$. (Under the assumption of $\lim_{c_t\to 0} u(c_t) = -\infty$, which implies the Inada condition that $\lim_{c_t\to 0} u'(c_t) = \infty$, consumption must be positive.[21])

The Lagrangian can be written as:

$$\begin{aligned}
L = -\bar{w} &+ \mu_0\left[\sum_{t=1}^{n} u(c_t)\gamma_t(\lambda) - V(\lambda) - \bar{u}\right] + \mu_1\left[\sum_{t=1}^{n} u(c_t)(\gamma_t^H - \gamma_t^L) - V'(\lambda)\right]\\
&+ \sum_{t=1}^{n} \beta_t\gamma_t(\lambda)(\bar{w} - c_t) + \sum_{t=1}^{n} \alpha_t\gamma_t(\lambda)c_t.
\end{aligned}$$

[21] See MacLeod (2003), pp. 219.

The first-order conditions are:

$$\frac{\partial L}{\partial \bar{w}} = -1 + \sum_{t=1}^{n} \beta_t \gamma_t(\lambda) = 0 \quad \text{i.e.,} \quad \sum_{t=1}^{n} \beta_t \gamma_t(\lambda) = 1, \tag{2.8}$$

$$\frac{\partial L}{\partial c_t} = \mu_0 u'(c_t)\gamma_t(\lambda) + \mu_1 u'(c_t)(\gamma_t^H - \gamma_t^L) - \beta_t \gamma_t(\lambda) + \alpha_t \gamma_t(\lambda) = 0, \tag{2.9}$$

and $$\frac{\partial L}{\partial \mu_0} \geq 0,\ \frac{\partial L}{\partial \mu_1} \geq 0,\ \frac{\partial L}{\partial \beta_t} \geq 0,\ \frac{\partial L}{\partial \alpha_t} \geq 0,$$

$$\mu_0, \mu_1, \beta_t, \alpha_t \geq 0,$$

$$\mu_0 \frac{\partial L}{\partial \mu_0} = 0,\ \mu_1 \frac{\partial L}{\partial \mu_1} = 0,\quad \beta_t \frac{\partial L}{\partial \beta_t} = 0,\ \alpha_t \frac{\partial L}{\partial \alpha_t} = 0,\ \forall t.$$

From (2.8), for some t signal(s) it must be that $\beta_t > 0$, which implies $\bar{w} = c_t$. However, this cannot be true for all t, otherwise there will be no incentive to exert any effort.

Denote $\bar{t}$ to be any signal for which there will be no money burning, i.e. $\bar{w} = c_{\bar{t}} > 0$. By complementary slackness, $\alpha_{\bar{t}} = 0$ for $\bar{t}$. Then (2.9) can be simplified as:

[No money burning] $$\mu_0 + \mu_1 \frac{\gamma_{\bar{t}}^H - \gamma_{\bar{t}}^L}{\gamma_{\bar{t}}(\lambda)} = \frac{\beta_{\bar{t}}}{u'(c_{\bar{t}})} = \frac{\beta_{\bar{t}}}{u'(\bar{w})} > 0. \tag{2.10}$$

Now consider all t such that $\bar{w} > c_t$, which implies $\beta_t = 0$. Then (2.9) can be written as:

$$\mu_0 + \mu_1 \frac{\gamma_t^H - \gamma_t^L}{\gamma_t(\lambda)} = -\frac{\alpha_t}{u'(c_t)}. \tag{2.11}$$

Denote $\underline{t}$ to be any signal such that consumption is zero, i.e. $c_{\underline{t}} = 0$, a case of full money burning where $\alpha_{\underline{t}} \geq 0$; and t' to be any signal such that there is only partial money burning, i.e. $\bar{w} > c_{t'} > 0$ where $\alpha_{t'} = 0$.

By Lemma 2, $\underline{t}, t' \in \mathcal{T}^-$. Then divide the case pertaining to (2.11) into:

$$\text{[Partial money burning]} \qquad \mu_0 + \mu_1 \frac{\gamma^H_{t'} - \gamma^L_{t'}}{\gamma_{t'}(\lambda)} = 0, \tag{2.12}$$

$$\text{[Full money burning]} \qquad \mu_0 + \mu_1 \frac{\gamma^H_{\underline{t}} - \gamma^L_{\underline{t}}}{\gamma_{\underline{t}}(\lambda)} = -\frac{\alpha_{\underline{t}}}{u'(c_{\underline{t}})} = -\frac{\alpha_{\underline{t}}}{u'(0)} \leq 0. \tag{2.13}$$

The last inequality of (2.13) holds true since the marginal utility of zero consumption is bounded without the Inada condition.

Now from (2.10), (2.12) and (2.13) we have:

$$\mu_0 + \mu_1 \frac{\gamma^H_{\bar{t}} - \gamma^L_{\bar{t}}}{\gamma_{\bar{t}}(\lambda)} > 0 = \mu_0 + \mu_1 \frac{\gamma^H_{t'} - \gamma^L_{t'}}{\gamma_{t'}(\lambda)} \geq \mu_0 + \mu_1 \frac{\gamma^H_{\underline{t}} - \gamma^L_{\underline{t}}}{\gamma_{\underline{t}}(\lambda)}. \tag{2.14}$$

Since $\mu_1 \geq 0$, from the first inequality it is clear that $\mu_1 \neq 0$ and $\mu_1 > 0$. By Lemma 1, the above inequalities imply that:

$$\bar{t} > t' \geq \underline{t}.$$

Further, the equality in the above chain in (2.14) holds for only a unique t'. At t', $\mu_0 = -\mu_1 \frac{\gamma^H_{t'} - \gamma^L_{t'}}{\gamma_{t'}(\lambda)} > 0$.

If $t' = \underline{t}$, the optimal consumption is $\bar{w} - b$ or higher and there might not be any t with full money burning (i.e., in effect, then, $\underline{t}$ will fail to exist). Note also that $\underline{t}$ $(\neq t')$ exists if and only if $\alpha_{\underline{t}} > 0$.

Now we solve the principal's problem. Since $\mu_0, \mu_1 > 0$, the IC and PC constraints hold with equality. Let[22]

$$c_t = \begin{cases} 0, & \text{if } \; t < t' \\ \bar{w} - b, & \text{if } \; t = t' \\ \bar{w}, & \text{if } \; t > t'. \end{cases} \tag{2.15}$$

[22] If $t' = 1$, the contract degenerates to MacLeod's contract.

Then we have:

$$\begin{aligned}
&\text{[IC]} \qquad u(0)\sum_{t=1}^{t'-1}(\gamma_t^H-\gamma_t^L)+u(\bar{w}-b)(\gamma_{t'}^H-\gamma_{t'}^L)+u(\bar{w})\sum_{t=t'+1}^{n}(\gamma_t^H-\gamma_t^L)=V'(\lambda),\\
&\text{[PC]} \qquad u(0)\sum_{t=1}^{t'-1}\gamma_t(\lambda)+u(\bar{w}-b)\gamma_{t'}(\lambda)+u(\bar{w})\sum_{t=t'+1}^{n}\gamma_t(\lambda)=V(\lambda)+\bar{u}.
\end{aligned} \tag{2.16}$$

Then we can explicitly solve:[23]

$$u(\bar{w})=\frac{V'(\lambda)\gamma_{t'}(\lambda)-[V(\lambda)+\bar{u}](\gamma_{t'}^H-\gamma_{t'}^L)+u(0)\left[(\gamma_{t'}^H-\gamma_{t'}^L)\sum_{t=1}^{t'-1}\gamma_t(\lambda)-\gamma_{t'}(\lambda)\sum_{t=1}^{t'-1}(\gamma_t^H-\gamma_t^L)\right]}{\gamma_{t'}(\lambda)\sum_{t=t'+1}^{n}(\gamma_t^H-\gamma_t^L)-(\gamma_{t'}^H-\gamma_{t'}^L)\sum_{t=t'+1}^{n}\gamma_t(\lambda)},$$

$$u(\bar{w}-b)=$$

$$\frac{[V(\lambda)+\bar{u}]\sum_{t'+1}^{n}(\gamma_t^H-\gamma_t^L)-V'(\lambda)\sum_{t'+1}^{n}\gamma_t(\lambda)-u(0)\left[\left(\sum_{t=1}^{t'-1}\gamma_t(\lambda)\right)\left(\sum_{t'+1}^{n}(\gamma_t^H-\gamma_t^L)\right)-\left(\sum_{t'+1}^{n}\gamma_t(\lambda)\right)\left(\sum_{t=1}^{t'-1}(\gamma_t^H-\gamma_t^L)\right)\right]}{\gamma_{t'}(\lambda)\sum_{t=t'+1}^{n}(\gamma_t^H-\gamma_t^L)-(\gamma_{t'}^H-\gamma_{t'}^L)\sum_{t=t'+1}^{n}\gamma_t(\lambda)}$$

Assuming that $u(0)=0$, then it is easy to show that $0<b\leq\bar{w}$, if and only if:

$$\frac{\sum_{t=t'}^{n}(\gamma_t^H-\gamma_t^L)}{\sum_{t=t'}^{n}\gamma_t(\lambda)}<\frac{V'(\lambda)}{V(\lambda)+\bar{u}}\leq\frac{\sum_{t=t'+1}^{n}(\gamma_t^H-\gamma_t^L)}{\sum_{t=t'+1}^{n}\gamma_t(\lambda)}. \tag{2.17}$$

Let

$$\phi(\tilde{t})=\frac{\sum_{t=\tilde{t}}^{n}(\gamma_t^H-\gamma_t^L)}{\sum_{t=\tilde{t}}^{n}\gamma_t(\lambda)},\ \text{where } \tilde{t}\in\mathcal{T}^-\cup\{t=K+1\}.$$

Note that $\phi(1)=0$ and it is easy verify that $\phi(\tilde{t})$ is increasing in $\tilde{t}$. Given that $\frac{V'(\lambda)}{V(\lambda)+\bar{u}}>0$, the left-hand inequality in (2.17) will be true for at least one t'. Therefore, if $\frac{V'(\lambda)}{V(\lambda)+\bar{u}}\leq\phi(K+1)$, then $t'\geq 1$ is uniquely determined by the inequality (2.17). This formally establishes (2.15) to be the optimal contract for the principal, completing the proof for part (i).

[23] If the utility function is linear, the solution is as follows:

$$b=\frac{V'(\lambda)\gamma_G(\lambda)-(V(\lambda)+\bar{u})(\gamma_G^H-\gamma_G^L)}{(\gamma_G^H-\gamma_G^L)\gamma_{t'}(\lambda)-(\gamma_{t'}^H-\gamma_{t'}^L)\gamma_G(\lambda)},\qquad \bar{w}=\frac{V'(\lambda)\gamma_{t'}(\lambda)-(V(\lambda)+\bar{u})(\gamma_{t'}^H-\gamma_{t'}^L)}{(\gamma_G^H-\gamma_G^L)\gamma_{t'}(\lambda)-(\gamma_{t'}^H-\gamma_{t'}^L)\gamma_G(\lambda)},$$

where $\gamma_G^H=\sum_{t=t'}^{n}\gamma_t^H$, and γ_G^L and $\gamma_G(\lambda)$ are defined accordingly.

If, on the other hand, $\frac{V'(\lambda)}{V(\lambda)+\bar{u}} > \phi(K+1)$, then no $t \in \mathcal{T}^-$ will qualify as t'. However, for the time being let us force $t' = K$ in solving b and $\bar{w}$ as done above. This means that besides all money $\bar{w}$ burnt for $t = 1, 2, \ldots, K{-}1$, money burning given by our artificially constructed solution b at $t = K$ has to exceed the principal's budget $\bar{w}$ in order to make both the IC and PC constraints bind (as in (2.16)). However, this is not possible as b must not exceed $\bar{w}$. This means the possibility of partial money burning cannot be sustained in the optimal contract, so it must be either a case of full money burning or no money burning:

$$c_t = \begin{cases} 0 & \text{if } \ t \in \mathcal{T}^-, \\ \bar{w} & \text{if } \ t \in \mathcal{T}^+. \end{cases}$$

Now using the same $\bar{w}$ as artificially constructed above and lowering b to equal $\bar{w}$ in (2.16), we will see that this last all-or-nothing money burning contract will create a slack in PC but will fail IC. So to restore the IC, $\bar{w}$ must be raised until IC binds and the slack in PC increases further, as follows:

$$\text{[new IC]} \qquad \sum_{t=K+1}^{n} u(\bar{w})(\gamma_t^H - \gamma_t^L) - V'(\lambda) = 0,$$

$$\text{[new PC]} \qquad \sum_{t=K+1}^{n} u(\bar{w})\gamma_t(\lambda) - V(\lambda) - \bar{u} > 0.$$

This new $\bar{w}$ with $u(\bar{w}) = \frac{V'(\lambda)}{\sum_{t=K+1}^{n}(\gamma_t^H - \gamma_t^L)}$ is the minimal budget that will implement λ. This completes the proof of part (ii). **Q.E.D.**

That is, below some threshold performance signal all money will be burnt and above the threshold the agent receives the full reward. Given monotone likelihood ratio condition, the principal gradually works up the signal ladder to burn money starting from the signal that is most indicative of lack of effort until the agent's marginal effort incentive condition binds (for the particular λ). Further, money burning happens only when a signal is more likely following the project's failure than if it is a success. If instead money is burnt for some t such that $\gamma_t^H > \gamma_t^L$, this would clearly dampen agent's effort incentive: a higher chance of success (from incremental effort) and thereby an improved chance of the signal t realizing means a higher probability of a strictly lower reward.

MacLeod (2003) gave us a fresh direction in a departure from classical contract theory models. He had argued that the constraints of subjective evaluations could be damaging as the principal will have much less freedom in incentives design. His extreme wage compression hypothesis embodies this idea. What was not very clear though is exactly what aspects of MacLeod's model were critical to the extreme wage compression result. Three assumptions distinguished his analysis – agent risk aversion, the monotone likelihood ratio property of signals, and agent ruin near zero consumption. Risk aversion suggests that perhaps the principal should not lump the penalty to just the lowest signal, as an appropriately small transfer of penalty from the lowest to the second-lowest signal should improve the agent's expected utility by lowering dispersion in consumption which ultimately would have been beneficial for the principal. So by assuming risk aversion, perhaps, MacLeod made it more difficult to derive extreme compression. If one were to drop risk aversion in favor of risk neutrality, there is no sound economic reason why the extreme wage compression result should be any harder to establish; on the contrary, the case for extreme compression should gain an additional ground. This leaves us with the remaining two explanations – ruin near zero consumption and the $MLRC$ assumption. The ruin assumption was forcing the principal not to push the agent's consumption down to zero. But still it does not explain why the principal should not go up the information ladder to burn money for signals higher than the lowest signal. We will argue that it is precisely here that $MLRC$ comes to play an important role. A first pass at why money burning does not happen at signals $t > 1$ would be as follows: given that at signal $t = 1$ money burning is only partial, for any money burning at $t = 2$ (or higher), part of it can be transferred to $t = 1$ signal and by doing so the principal would rely on the greater informativeness of $t = 1$ signal than $t = 2$ signal ($MLRC$). But then risk aversion counters this manoeuver, as noted earlier. To fully understand how $MLRC$ wins over the opposite force of risk aversion, we need a more careful argument, which we detail below.

Starting with MacLeod's optimal mechanism, let us see how any attempt to extend money burning beyond the lowest signal is counterproductive for both the agent and the principal. For our argument, we relax MacLeod's requirement that the consumption at the lowest signal must be strictly positive, and instead let us assume that at $t = 1$ the agent consumes $w \geq 0$. The agent's expected utility and marginal effort incentives are then given by, respectively,

$$\gamma_1(\lambda)u(w) + \gamma_2(\lambda)u(w+b) + \sum_{t=3}^{n} \gamma_t(\lambda)u(w+b) - V(\lambda) - \bar{u}\,,$$

$$(\gamma_1^H - \gamma_1^L)u(w) + (\gamma_2^H - \gamma_2^L)u(w+b) + \sum_{t=3}^{n}(\gamma_t^H - \gamma_t^L)u(w+b) - V'(\lambda)\,.$$

Now increase consumption at $t = 1$ by $\delta_1 > 0$ and lower consumption at $t = 2$ by $\delta_2 > 0$ suitably (while keeping consumption at $t > 2$ the same) such that the agent's expected utility, hence PC, does not get hurt:

$$\gamma_1(\lambda)[u(w+\delta_1) - u(w)] + \gamma_2(\lambda)[u(w+b-\delta_2) - u(w+b)] \geq 0. \quad (2.18)$$

This last adjustment in the rewards hurts the agent's effort incentive

(IC), if

$$(\gamma_1^H - \gamma_1^L)[u(w) - u(w+\delta_1)] + (\gamma_2^H - \gamma_2^L)[u(w+b) - u(w+b-\delta_2)] \ > \ 0,$$

i.e.,
$$\frac{u(w+b) - u(w+b-\delta_2)}{u(w) - u(w+\delta_1)} \ < \ \frac{\gamma_1^H - \gamma_1^L}{\gamma_2^H - \gamma_2^L}. \tag{2.19}$$

From (2.18) it follows that

$$\frac{u(w+b) - u(w+b-\delta_2)}{u(w) - u(w+\delta_1)} \ \leq \ \frac{\gamma_1(\lambda)}{\gamma_2(\lambda)} \ \underbrace{<}_{\text{(by Lemma 1)}} \ \frac{\gamma_1^H - \gamma_1^L}{\gamma_2^H - \gamma_2^L},$$

thus verifying (2.19).

We can now see that if the principal tries to shift money burning from the lowest signal to any other signal while maintaining the same maximal reward, he is going to damage the agent's effort incentive, which ultimately will hurt his own objective. This negative implication is all driven by the $MLRC$ property and holds irrespective of whether the agent is risk averse or risk neutral. Another way to view the above is to say that, if the principal had a way to lower money burning at any $t > 1$ (from, say, $w + b - \delta_2$) and replace it with increased money burning at $t = 1$ (say from a strictly positive consumption), he would do so. This implies, positive money burning at any $t > 1$ cannot be an optimal response in MacLeod's model so long as consumption at $t = 1$ is strictly positive. But then the Inada assumption in MacLeod's model forces the consumption at $t = 1$ to be strictly positive. This rules out positive money burning at any signal other than the worst signal. In contrast, in our formulation, dropping of the Inada condition implies that the principal can already set money burning to be maximal (or full) at $t = 1$, if he wishes so. So the only way our principal can manoeuver incentives further, in terms of shifting money burning around, is to go up the signal order, starting from $t = 1$. If our principal does not extend money burning beyond the lowest signal, then to provide marginal (effort) incentive to the agent the principal will have to increase the overall reward from $w + b$ to some $w + b + \Delta$ and then set full money burning at $b + \Delta$ at $t = 1$, but this is more costly than the alternative of extending money burning beyond $t = 1$.

To summarize, so long as $MLRC$ is viewed as a very basic assumption in any moral hazard model of contracting, the main difference between extreme and moderate wage compression comes from the assumption of Inada condition (or its absence). On the face of it, Inada condition is usually not considered such a serious imposition in most models (such as growth model) and its main purpose is to ensure interior solutions. However, in our contracting environment under *spe*, the condition turns out to be both necessary and sufficient for extreme wage compression. And if the agent possesses some wealth of his own or sometimes a minimal positive consumption is guaranteed by the state through the social welfare program, then we recover moderate wage compression even with the Inada condition.

■ Comparison of two-period model with the static model. Our modification of MacLeod's model should also be useful for comparison with the finite repeated game models. Without interim feedback and interim money burning (that are shown to be optimal in Propositions 3-**??**), the repeated efforts model can be seen in the same way as a static game: in the most discriminating case of Proposition 5 (i.e., part **(II)**), the signals $\{\sigma_H \sigma_L, \sigma_L \sigma_H\}$ can be viewed as the cutoff t' with $\bar{w} = \frac{2c}{p_1 - p_0}$ and $b = \frac{c}{p_1 - p_0}$ in part (i) of Proposition 6; part **(I)** of Proposition 5 parallels part (ii) of Proposition 6; and part **(III)** of Proposition 5 parallels the extreme wage compression result of MacLeod (Proposition 6-MacLeod (2003)). It may be noted that the pay for performance in our two-period model is part of the more general 2-fold wage compression incentives in the static model.

2.5 Robustness check: Impact of agent information

Chan and Zheng (2011) (in short, CZ) study a principal-agent contracting problem similar to ours but they introduce the possibility that the agent might have some information about his own performance.[24] By incorporating agent information that is correlated with the principal's signal, the optimal money burning mechanism is shown by the authors to reflect the *pay for performance* principle. In a two-period example, they show that money burning should be more substantial when the performance is declining over time (signal profile "high-low") than if it is improving (signal profile "low-high").

CZ's model differs from ours in the principal's objective similar to the difference between our model and Fuchs (2007). CZ assume that the agent might be asked by the principal to pay upfront a lump-sum amount before period 1 which the agent may forfeit following low performance. This violates the agent's limited liability and converts the principal's profit-maximization problem into maximization of the social surplus. Thus whether the pay-for-performance result will hold when the agent has additional information and the principal maximizes profits and the agent is subjected to limited liability, remains to be seen. In this section, we do this verification.

CZ derive the following optimal incentive scheme when $\rho > 0$:

$$W = \frac{1+\rho(1-p_1)}{1-p_1}\frac{c}{p_1-p_0}$$

$$z^{HH} = 0, \quad z^{HL} = \rho\frac{c}{p_1-p_0}, \quad z^{LH} = 0, \quad z^{LL} = \frac{1+\rho(1-p_1)}{1-p_1}\frac{c}{p_1-p_0},$$

[24] CZ's analysis is for finite T period repeated games with specific result for $T = 2$ relevant for comparison with our results.

where $\rho \in (0, 1)$ is the correlation between the agent's and the principal's signals.[25]

Now let us introduce agent private information to our two-period model of section 3. Let the agent receive a binary signal $s_1 \in \{s_G, s_B\}$ after period one. This private signal, when realized, together with the private effort chosen, will form the agent's first-period posterior belief of the principal's first-period signal: $\Pr[\sigma_1 = \sigma_H \mid (e_1, s_1)]$.[26] The posterior is defined as follows:[27]

$$\begin{aligned}
q^{1G} &= \Pr[\sigma_1 = \sigma_H \mid (1, s_G)] = p_1 + \rho(1 - p_1) \\
q^{1B} &= \Pr[\sigma_1 = \sigma_H \mid (1, s_B)] = p_1 - \rho(1 - p_1) \\
q^{0G} &= \Pr[\sigma_1 = \sigma_H \mid (0, s_G)] = p_0 + \rho(1 - p_0) \\
q^{0B} &= \Pr[\sigma_1 = \sigma_H \mid (0, s_B)] = p_0 - \rho(1 - p_0).
\end{aligned}$$

We have the following implications:

$$q^{1G} > p_1 > q^{1B}, \quad q^{0G} > p_0 > q^{0B}.$$

To simplify the analysis, we further impose the following assumption:

ASSUMPTION 1 (**Limited correlation between signals**). *Let $q^{1B} > q^{0G}$, or equivalently, $0 < \rho < \frac{p_1 - p_0}{2 - p_1 - p_0}$.*

What this assumption means is that while the agent's self-evaluation possibly reflects also the principal's evaluation, the information that the agent exerted effort means a higher likelihood of principal receiving a high signal even when accompanied by a bad signal on the agent's side

[25] Since the intertemporal discounting is assumed away in this work, the corresponding case in CZ is when $\rho > 1 - \delta$. We report only this case.

[26] Communication or renegotiation between the agent and the principal after period 1 is not considered.

[27] Recall that p_1 and p_0 refer to the prior probabilities that the principal will receive a high performance signal with and without effort exerted in that particular period.

compared to when the agent does not exert effort but observes a good signal.

With the introduction of agent information the time line of the game changes as follows:

1. *At time zero, a contract* $(W, \mathbf{z}(\Sigma))$ *is signed between parties;*

2. *In period* 1*, the agent decides whether to exert effort or shirk; at the end of the period, the agent and the principal each privately observe a performance signal;*

3. *In period* 2*, the agent chooses effort once more based on first-period performance; at the end of the period, the principal reports the performance signal profile* $\{\sigma_1\sigma_2\}$ *and makes the payment accordingly.*

The principal's objective is to minimize the reward costs of implementing full efforts by the agent over two rounds, $(1, 1)$. Using backward induction, start from the second period. The information based on which the agent makes the second-period effort decision is defined by first-period effort and signal pair (e_1, s_1). From each such pair, there is a proper subgame for the agent as illustrated in Fig. 2.2. To implement the

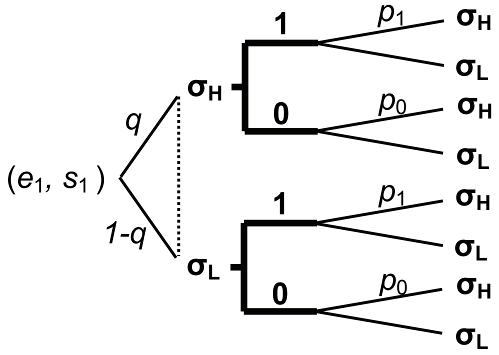

Figure 2.2: Subgame following (e_1, s_1)

second-period effort, the rewards (and money burning amounts) should be designed such that upon first-period effort, i.e., at either information of $(1, s_G)$ or $(1, s_B)$, the agent should have higher expected continuation value (or lower expected cost) from choosing $e_2 = 1$ than $e_2 = 0$. This will give us the incentive compatibility conditions for exerting effort in the second period. Then, going back to the first period we need to make sure that the expected value for the agent is the highest (or expected cost is lowest) for the effort profile $(1, 1)$, among $\{(1, 1), (1, 0), (0, 1), (0, 0)\}$. This will lead to the incentive compatibility conditions for full efforts for two periods. Then imposing the participation constraints in equilibrium, as well as the non-negativity constraints, we can write the principal's cost-minimization problem formally (see Appendix). Solving this problem yields the optimal money burning mechanism as shown in Table 2.2.[28,29]

(i) $p_1 + p_0 > 1$	**(ii)** $p_1 + p_0 = 1$	**(iii)** $p_1 + q^{1G} \leq 1$
$W = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$W = \frac{2c}{p_1-p_0}$	$W = \frac{1+\rho(1-p_1)}{1-p_1}\frac{c}{p_1-p_0}$
$z^{HH} = 0$	$z^{HH} = 0$	$z^{HH} = 0$
$z^{HL} = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$z^{HL} = \frac{c}{p_1-p_0}$	$z^{HL} = \rho\frac{c}{p_1-p_0}$
$z^{LH} = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$z^{LH} = \frac{c}{p_1-p_0}$	$z^{LH} = 0$
$z^{LL} = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$z^{LL} = \frac{2c}{p_1-p_0}$	$z^{LL} = \frac{1+\rho(1-p_1)}{1-p_1}\frac{c}{p_1-p_0}$

Table 2.2: Optimal money burning mechanism with agent's information

Table 2.2 illustrates three different cases. In cases (i) and (ii), money burning is necessary whenever there is a low performance signal; also, the money burnt following "high-low" signals is same as the one following "low-high" signals, i.e., the order of signals does not matter given the same number of low signals – a result in contrast with Chan and Zheng

[28] Note that full efforts need not always be implementable, so the cases in Table 2.2 are mutually exclusive but not necessarily exhaustive.

[29] The detailed derivations are provided in the Appendix.

(2011). In case (iii), money burnings upon $\sigma_H\sigma_L$ and $\sigma_L\sigma_H$ are asymmetric, the former positive and the latter equal to 0 – a result similar to CZ's result. Given these observations, next we address two questions: first, how do changes in agent's private information impact on the optimal mechanism as derived in our Proposition 5; second, with agent information added, how does the principal's implementation cost compare with the one in CZ (where the principal effectively minimizes expected money burning).

When $p_1 + p_0 > 1$, under the mechanism derived in the absence of agent information (see case (I) in Proposition 5), the agent strictly prefers exerting effort to shirking in the second period, i.e., the incentive compatibility condition for choosing $(1, 1)$ over $(1, 0)$ is slack. Now with the introduction of agent's self-evaluation, he knows more information about the principal's first-period signal, and additional incentive compatibility conditions are required for implementing second-period effort. However, no matter his information is good (s_G) or bad (s_B), it turns out that the original mechanism still makes him strictly better off exerting effort rather than shirking. Therefore, the private signal of the agent does not change the original optimal mechanism in this case. When $p_1 + p_0 = 1$, the additional IC conditions imposed for second period agent effort happen to also bind under the mechanism proposed in case (II) of Proposition 5, so that agent information does not alter the previous optimal mechanism in this case too. Therefore, if $p_1 + p_0 \geq 1$, which implies the principal's signal reflects, on average, agent's effort,[30] the principal behaves as if there is no agent information.

When $p_1 + q^{1G} \leq 1$, which implies $p_1 < 1/2$ i.e. that the principal's signal is not as informative, introduction of agent's private information does alter the optimal mechanism. In particular, for all $p_1 \leq 1/2$, the orig-

[30] Note that $p_1 + p_0 \geq 1$ implies $p_1 > 1/2$ since $p_1 > p_0$.

inal mechanism in Proposition 5 suggests burning money only upon the worst signal realization $\sigma_L\sigma_L$, and the incentive compatibility conditions for full efforts $(1,1)$ over both $(1,0)$ and $(0,1)$ are binding. Now with the agent's private signal, for example a good signal s_G, the probability that the agent believes σ_H is received by the principal is higher than its prior (i.e., $q^{1G} > p_1$); as a result, the incentive provided by the previous mechanism without the agent's information, $z^{HL} = 0$, becomes insufficient: the agent would now strictly prefer shirking in the second period. Therefore, any positive chance of getting s_G necessitates positive money burning upon $\sigma_H\sigma_L$ signals, discouraging the agent from shirking after any sign of first-period success. Furthermore, with z^{HL} positive, the original value of z^{LL} has to be pushed up too in order to maintain the incentives provided by $z^{LL} - z^{HL}$, otherwise the agent would strictly prefer shirking in the first period. Therefore, the principal's cost is pushed up accordingly. Below we summarize these observations.

PROPOSITION 7 (**Impact of agent's information**). *Suppose Assumption 1 holds.*

(i) If $p_1 + p_0 \geq 1$, the optimal mechanism is not affected by the agent's private information and may exhibit the ***pay for performance*** *principle (cases* **(i)** *and* **(ii)** *in Table 2.2).*

(ii) If $p_1 + q^{1G} \leq 1$, the optimal mechanism depends on the fact that the agent also has some private information about his own performance: the principal has to incur a higher implementation cost to induce full efforts with more frequent money burning than if the agent were ignorant. Again, the optimal mechanism exhibits the ***pay for performance*** *principle (case* **(iii)** *in Table 2.2) that depends on the specifics of time structure of progress similar to Chan and Zheng's (2011) result.*

Now let us turn to a comparison with CZ's optimal mechanism reported earlier. It can be seen that the mechanism is identical to that in case (iii) of ours: both result in higher implementation costs (than the mechanism derived without agent information), and punish the agent asymmetrically upon $\sigma_H \sigma_L$ and $\sigma_L \sigma_H$. However, in some circumstances, say as in cases (i) and (ii), our objective of minimizing principal's cost results in a lower cost and more frequent but symmetric money burning than CZ's mechanism, which is derived from the objective of minimizing expected money burning. When $p_1 + p_0 \geq 1$, as noted in Proposition 7, the principal ignores agent's information and burns money whenever a low signal is observed. By punishing the agent more frequently, our mechanism requires less amount being burnt per incident, at the expense of more total expected money burning. However, to minimize expected money burning, as CZ's principal seems to target, would require invoking the "reuse of punishment" principle, i.e., shift money burning from some circumstances to the worst signal profile. This practice pushes up the principal's cost. When $p_1 + q^{1G} \leq 1$, it might not be optimal to punish the agent whenever a low signal is observed, which is more likely to be a consequence of bad luck rather than zero effort. In this case, minimizing principal's cost yields identical solution to the one when the principal minimizes the social cost (due to the deadweight loss of money burning).

PROPOSITION 8 (**Comparison with Chan $\&$ Zheng's mechanism**). *Suppose Assumption 1 holds.*

(i) If $p_1 + p_0 \geq 1$, the objective of minimizing principal's cost leads to a mechanism with more frequent money burning that is also symmetric, i.e. money burning amount is the same whether for improving or declining performance, in contrast to Chan and Zheng's result. Overall, our mechanism involves a lower implementation cost for the principal than their money burning mechanism.

(ii) If $p_1 + q^{1G} \leq 1$, our mechanism and Chan and Zheng's mechanism yield identical recommendations.

2.6 Conclusion

Annual adjustments in base salaries and bonuses following performance review meetings with department managers or heads is a common occurrence in most organizations. Human resource departments mostly gather hard information about employee performance but when it comes to real decision making about pay adjustments, the words of someone with real authority often carry substantial weight. The earlier literature on subjective evaluation has argued that the performance based reward adjustments, especially when information about performance is soft, should be less common but more sharp that has come to be known as the (extreme) wage-compression hypothesis. This paper returns to this hypothesis.

We have argued that so long as employment relations are of finite duration (modelled as a two-period repeated efforts game or a static game), employees are risk averse or risk neutral but do not go broke at zero consumption, and employers seek to maximize profits rather than social efficiency, more sensitive performance pay or a threshold based wage compression is typically the optimal strategy for employers. Thus, the additional optimal contracting possibilities tend to be relatively more discriminating as in Levin (2003), as opposed to the single-incidence based punishment schemes of MacLeod (2003), and Fuchs (2007).

While this paper is definitely not about how to endogenize money burning, which we accept as the one shortcoming of our analysis, it is not difficult to contemplate money burning in practice as organizations do rely on fixed salary/bonus pools to reward their employees. Analyzing single-agent incentive schemes against such a backdrop is a plausible step forward to understand better organizational practices. This work should be seen as complementary to MacLeod (2003), Levin (2003), and Fuchs (2007) on subjective evaluations, wage compression and performance pay.

CHAPTER 3

Subjective Performance Evaluation and Perils of Favoritism

3.1 Introduction

Favoritism in organizations, if not rampant, is not uncommon. In team settings treating identical agents differently can have economic efficiency. This is shown by Winter (2004) in a multi-agent, moral hazard problem involving multiple tasks where agents perform one task each and the team project succeeds for sure if all tasks are completed successfully. To minimize the incentive costs, the principal rewards the agents differentially if the overall project succeeds but otherwise the agents receive the same zero reward. In Winter's analysis team performance is verifiable but individual efforts are not.[1]

We consider a static, one-period team problem where team performance is not immediately known, nor are individual efforts. Instead, the principal privately observes only a signal, high or low, of the team's collective performance. The agents either exert effort or shirk, and with an additional effort the positive signal of performance is more likely. Since the signal is not verifiable and hence not contractible, the performance evaluation and eventual distribution of rewards are essentially *subjective*. We first show that, like in Winter, the principal would discriminate be-

[1] In an experimental work on Winter's model, Sebastian, Sebastian, and Zultan (2010) show that unequal rewards can potentially increase productivity by facilitating coordination.

tween ex-ante identical agents by promising differential positive rewards if he observes a positive signal of performance, but otherwise give zero rewards. Zero rewards necessitate *money burning*. Then we ask the following question: how should an organization deal with money burning?

It is well known that for *subjective performance evaluation*, in short *spe*, money burning is a necessary evil especially in static environments. Organizations rarely, if at all, engage in explicit money burning. After all, there is always the question of accountability: how would the manager justify money burning? Economic reasons cannot persuade outsiders, be they tax payers in the case of governmental organizations or shareholders for private firms, about the need for clear *wastage*. One explanation put forward by MacLeod (2003) (who studies a one-period principal-agent problem with agent moral hazard and *spe*) is that the principal can transfer the 'burnt money' to a third party or equivalently hire an extra agent. In concrete terms what this means is that the principal will have employees who, despite not being central to the team's core activity in question, could become the lucky beneficiaries. Or a more plausible scenario is one where the principal openly identifies his favorite employee (or it becomes common knowledge who is the favorite employee) who receives higher rewards in each contingency and more so when the team performs poorly.[2] But such *favoritism* or *organizational bias* will lead to another problem, that of *sabotage*, that seriously undermines an organization's attempt at incentive provision (as suggested by Winter) or avoidance of wastage in money burning (as hinted by MacLeod). And worse still, sometimes there might not be any *sabotage-proof mechanism* (when money burning is replaced by a balanced budget mechanism) by which all key members of a team can be induced to put in pro-

[2] Rasmusen (1987) studies moral hazard problem in risk-averse teams and proposes a "massacre" contract: all but one randomly selected agent are punished and the lucky one gets rewarded upon bad team performance. His work is not in the *spe* context nor does he consider the possibility of sabotage or its implications.

ductive efforts rather than shirk. Ultimately, an organization may have to choose between two evils – *money burning* and *back-stabbing and scheming* within its workforce. If money burning is not a choice, one could be left with only a scheming group. This is in addition to the familiar problem of collusion encountered in team settings as earlier noted by Eswaran and Kotwal (1984) (see the discussion at the end of section 4).

There are a number of important contributions on subjective performance evaluation but the paper that prompts us to go further on the issue of favoritism and its incentive role is MacLeod (2003), following the lead from Winter (2004).[3] We extend his (MacLeod's) principal-agent problem under *spe* to team problems as in Holmstrom (1982). Given that individual contributions are difficult to ascertain in a team game, the notion of *spe* should be adapted to team's collective performance. The principal obtains a private signal of team performance and gives rewards to team members based on the observed signal.[4] MacLeod had observed that (refer p. 222), "...the role of a good subjective evaluation system is not the elimination of socially wasteful conflict, but rather to find an optimal trade-off between the imposition of costs *ex post* on the relationship and the provision of performance incentives." In static, one-shot environ-

[3] Baker, Gibbons and Murphy (1994) consider the complementary role of subjective performance measure along with imperfect objective measures for employee incentives. Levin (2003) studied subjective performance evaluation in a repeated interaction principal-agent model, where the principal's assessment of the agent's performance in each period is private (see section IV). The author focused on simple truth-telling contracts in which the principal submits an accurate report of the performance signal observed. Levin doesn't concern with money burning, instead future payoff considerations incentivize the principal to report performance related information accurately. In our static environment on the other hand, only money burning can ensure truthful reporting.

[4] Giving rewards to a group of employees based on subjective impressions of the group's performance is not unusual. A close-to-home example is in academics where a department may be given its pool of annual bonus or salary increments by the Dean of the Faculty based on indicators about how the department as a collective unit is doing, where the indicators need not be objectively verifiable. For instance, the Dean may seek an external's opinion where there is quite a bit of subjectivity involved.

The CEO of a corporate firm may determine the rewards to the members of a research division on the basis of progress reports on hand about the team's activities. Assessment of performance related reports can be very subjective.

ments ex-post costs are due to money burning. Here we argue that how one models money burning (social waste), and in that regard adoption of a plausible formulation of one of MacLeod's suggestions (i.e., *budget balancing* via a third party), can have serious negative implications for incentives. Favoritism to avoid money burning may lead to instability even when there might be a prior compelling economic argument in favor of favoritism. Overall, our analysis suggests that there could be an upward bias in assessing the power of money burning for organizational incentives. This is not to deny other more congenial interpretations of money burning such as costs through conflicts in a repeated principal-agent relationship as in Levin (2003), but such models usually rely on the infinite game structure. For finitely repeated games (and ours is a one-shot game), money burning must be internalized through budget balancing for it to be credible as an organizational incentive mechanism. This will bring us back to the kind of team game modeled in this paper.

To the best of our knowledge, the only other paper to consider subjective performance evaluation in teams is by Rajan and Reichelstein (2006). They, like us, look at the problem of endogenizing money burning in a team moral hazard setting. Unlike in our case, which is the most significant difference, their principal obtains separate, independent signals of team members' performance. While the signals are private, the principal is still able to punish each agent according to his/her performance based on individual signals (budget balance within the team avoids money burning). The individualized punishment in Rajan and Reichelstein thus eases to a great extent the difficulties of incentive provision associated with free riding in teams. In fact, under suitable assumptions about the precision of signals the authors implement second-best and sometimes first-best efforts. Also, the authors do not consider sabotage. Thus, ours is a more traditional 'moral hazard in teams' problem with the focus on a coarse team-based subjective evaluation.[5] The sharp implications of favoritism and related sabotage incentives allow us to offer a much clearer perspective to the literature on subjective performance evaluation.[6]

The earlier works on favoritism in organizations (Prendergast and Topel, 1996; 1993; Prendergast, 1996) address the problem of supervisory bias in subjective evaluations of subordinates (due to a preference for exercising authority) and how organizations should respond to it by making agent incentives less or more high-powered. Money burning was never an issue in the above works. But relating favoritism to subjective evaluations remains a compelling account of organizations. In ours, subjectivity in evaluations gives rise to money burning and favoritism.

The rest of the paper is organized as follows. Subjective performance evaluation is introduced in section 2. In section 3, the principal's optimal money burning mechanism is derived when agents can engage only in productive efforts, followed by an analysis of sabotage in section 4. In section 5, we look more closely at the issue of fair treatment vs. effort implementation costs. Closing remarks are contained in section 6. Proofs appear in the Appendix.

[5] With individual signals and fixed total rewards budget agents can be incentivized by *relative performance evaluation*, whereas with coarse information a *joint performance evaluation* seems a natural choice. See Che and Yoo (2001) for this distinction.

[6] Sabotage in contests is a well-studied topic (see, for example, Konrad (2000)). For team problems, Eswaran and Kotwal (1984) had pointed out the possibility of a principal colluding with an agent to induce him to lower his effort when budget balance is broken by giving the entire team output to the principal (or a third party) for poor performance. We do not consider collusion between the third party and any of the core team members but instead give the third party, who is different from the principal, an active role in sabotage through costly effort and study whether this sabotage incentive can be carefully controlled by the principal and how it impacts on the effectiveness of subjective performance evaluation.

3.2 The model

A principal organizes a team of two or more workers (or agents), $i = 1, 2, 3$, in a joint project in a static one-shot game. The agents simultaneously decide whether to exert effort or shirk, $e_i \in \{0, 1\}$. Exerting one unit of effort will incur a cost c for the agent. The principal does not observe agents' actions but receives a private signal about team performance which is either high or low: $\sigma \in \{\sigma_H, \sigma_L\}$. The performance signal depends on the aggregate effort of the team as follows:

$$\Pr[\sigma_H \mid (e_i)] = \begin{cases} p_0, & \text{if} \quad \Sigma e_i = 0 \\ p_1, & \text{if} \quad \Sigma e_i = 1 \\ p_2, & \text{if} \quad \Sigma e_i \geq 2. \end{cases}$$

We thus assume that for the team project under consideration, only the best two efforts matter. Throughout this paper, the principal will be assumed to be interested in implementing two units of effort by involving two or at most three agents. In the case of three agents, the principal wants two specific agents to coordinate in exerting efforts as otherwise there will be too little or too much efforts.

Further, we impose the following assumptions on the technology:

[A1] $p_2 > p_1 > p_0$,

[A2] $p_2 - p_1 > p_1 - p_0$.

That is, the likelihood of a high performance signal is increasing in total efforts and the efforts are strategic complements.

Due to non-observability of agents' actions, the principal can reward the agents based only on team performance: agent i will receive r_i^H if σ_H is reported, and r_i^L if σ_L is reported; the rewards can be discriminating.

The total budget for the principal upon different signal realizations:

$$W^H = \sum_i r_i^H \qquad\qquad W^L = \sum_i r_i^L.$$

The time line of the game is as follows:

1. A self-enforcing implicit (or explict) contract involving performance related rewards is agreed to between the principal and the agents as a collective unit.
2. The agents simultaneously and privately choose their effort levels for team production.
3. The principal receives a signal of team performance.
4. The principal reports the performance signal and rewards the agents as he sees fit.

3.3 Money burning and favoritism

In this section, we consider alternative forms of implementation with principal inducing two units of efforts. The principal designs incentives to minimize implementation costs.

Under subjective evaluations, signals of team performance are not third-party verifiable and therefore the principal will report them truthfully provided he does not strictly gain from non-truthful reporting. The following implication is straightforward:

LEMMA 3. *The principal should commit to a fixed total budget W regardless of the realized signal.*

The intuition is simple. With the reward commitment being invariant to his reports, the principal should tell the truth. The result appeared in MacLeod (2003).

■ *Benchmark case*. Initially we consider only two agents. The principal's budget should satisfy:

$$W = r_1^H + r_2^H = r_1^L + r_2^L. \tag{3.1}$$

To implement $(e_1, e_2) = (1, 1)$, agent 1 should be rewarded more upon high signal than low signal as effort on his part implies marginally a higher probability of high signal. But then, given fixed budget, agent 2's reward should be less upon high signal than if it is a low signal. This implies agent 2 will have an incentive to shirk. The following result summarizes this dilemma.

PROPOSITION 9 (Non-existence of incentive compatible mechanism). *With two agents and private evaluation of team performance, there does not exist an incentive compatible reward scheme such that both agents will exert efforts with the rewards satisfying budget balance as in* (3.1)*.*

This negative result is well known in the team implementation literature (Holmstrom, 1982). To restore incentive compatibility, the budget balance must be broken with the principal *burning money* upon receiving a signal of poor team performance.

■ *Money burning mechanism*. Suppose now the principal destroys some surplus by paying to a third party upon receiving each performance signal.[7] The modified budget constraint is then:

$$W = r_1^H + r_2^H + z^H = r_1^L + r_2^L + z^L, \tag{3.2}$$

where z^H and z^L are the money burnings for high and low signals.

[7] MacLeod (2003) addresses the optimal contract design problem in a single agent, one period game with the possibility that part of the budget will be paid to a third party. In Bag and Qian (2013), we study the optimal money burning mechanism in a single agent, repeated efforts game.

As σ_H is a signal of good performance, there is no reason to burn any money, so set $z^H = 0$. Thus the incentives are augmented by money burning: (r_i^H, r_i^L, z^L). The induced effort game is as in G_1.

		Agent 2	
		0	1
Agent 1	0	$p_0 r_1^H + (1-p_0) r_1^L \, , \, p_0 r_2^H + (1-p_0) r_2^L$	$p_1 r_1^H + (1-p_1) r_1^L \, , \, p_1 r_2^H + (1-p_1) r_2^L - c$
	1	$p_1 r_1^H + (1-p_1) r_1^L - c \, , \, p_1 r_2^H + (1-p_1) r_2^L$	$p_2 r_1^H + (1-p_2) r_1^L - c \, , \, p_2 r_2^H + (1-p_2) r_2^L - c$

Figure 3.1: Two-agent game G_1

We next determine the incentive compatibility [IC] conditions such that $(e_1, e_2) = (1, 1)$ is the unique Nash equilibrium of G_1. Based on the [IC] conditions we then solve the principal's budget minimization problem.

Incentive compatibility constraints. For $(1, 1)$ to be a unique Nash equilibrium, we first require agent 1 to exert effort as his dominant strategy, and then agent 2 exerts effort as a Nash best response. The relevant conditions are as follows:

$$\begin{aligned} [\textbf{IC}_1] \qquad & r_1^H - r_1^L \geq \frac{c}{p_1 - p_0} \\ [\textbf{IC}_2] \qquad & r_2^H - r_2^L \geq \frac{c}{p_2 - p_1} \, . \end{aligned}$$

Participation constraints. The expected reward net of effort cost should be weakly greater than the opportunity cost of time, normalized to zero:

$$\begin{aligned} [\textbf{PC}_1] \qquad & p_2 r_1^H + (1-p_2) r_1^L - c \geq 0 \\ [\textbf{PC}_2] \qquad & p_2 r_2^H + (1-p_2) r_2^L - c \geq 0 \, . \end{aligned}$$

Finally, we solve the principal's cost minimization problem subject to

[ICs], [PCs], and the natural non-negativity conditions:

$$\min_{\{r_i^H, r_i^L, z^L\}} \quad W \qquad (\mathcal{P})$$

$$s.t. \qquad [\text{IC}_1], [\text{IC}_2], [\text{PC}_1], [\text{PC}_2] \text{ and } r_i^H \geq 0,\, r_i^L \geq 0,\, z^L \geq 0,$$

to derive the optimal mechanism.

PROPOSITION 10 (Money burning for unique Nash implementation). *For the two-agent team, the money burning mechanism implementing* $(e_1, e_2) = (1, 1)$ *in a unique Nash equilibrium at minimal cost to the principal is as follows:*

$$\begin{aligned} (\textit{Principal's cost}) \qquad & W^{MB} = \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1} \\ (\textit{Rewards}) \qquad & r_1^H = \frac{c}{p_1 - p_0}, \quad r_2^H = \frac{c}{p_2 - p_1}, \quad z^H = 0 \\ & r_1^L = 0\,, \quad r_2^L = 0\,, \quad z^L = \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1}. \end{aligned}$$

The above result extends the principal's optimal money burning mechanism of MacLeod (2003), and Bag and Qian (2013), from single-agent to a team setting. The optimal mechanism exhibits asymmetric treatments of identical agents. The basic argument follows Winter (2004): To uniquely implement two units of efforts, sufficient reward should be provided to one of the agents in order for him to have a dominant strategy to exert effort; then the other agent will exert effort with less reward, since the marginal benefit of his effort is higher given the guarantee of the first agent's effort due to complementarity of efforts. Our result thus extends Winter's (2004) *asymmetric treatment of identical agents* result for subjective performance evaluation.

■ *Adding a third agent to the team.* Money burning, despite its compelling economic intuition, is hard to justify in organizations. Taking a lead from MacLeod's (2003) suggestion, we probe how endogenizing money burning with the help of an additional non-functioning agent alters the incentives. We stress that while MacLeod's suggestion prompts us to look at an extra agent, eventually we empower this agent with an active role to influence team production by engaging in subversion of other team members' efforts.

With three agents, Lemma 3 still applies and the principal's budget constraint becomes:

$$W = r_1^H + r_2^H + r_3^H = r_1^L + r_2^L + r_3^L. \tag{3.3}$$

The normal form for the three-agent team is shown in the game G_2.

Since only two best efforts matter in generating team's performance signal, an agent would shirk if the other two choose to exert effort. Therefore, inducing three units of effort is neither possible, nor necessary:

LEMMA 4. *With three agents in the team, the principal is not able to incentivize all agents to engage in productive efforts at the same time.*

We will thus focus on inducing two agents to exert efforts, leaving the third to shirk. Let agent 3 be the one who is not expected to work, and the principal's objective is to implement $(e_1, e_2, e_3) = (1, 1, 0)$ at minimal cost.

In designing the optimal mechanism, the principal may also be concerned about other issues such as fairness among team members, strategic independence, etc. The trade-offs can be different with various types of implementations. We address these issues next.

■ *Unique Nash implementation.* Initially the third agent serves to simply absorb the burnt money, receiving it as a reward when team performs poorly but otherwise receives nothing. As one can see immediately from Proposition 10, the same incentives implement in a unique Nash equilibrium the effort profile $(1, 1, 0)$; agent 3's dominant strategy is to shirk.

		Agent 2	
		0	1
Agent 1	0	$p_0 r_1^H + (1-p_0) r_1^L\, , \, p_0 r_2^H + (1-p_0) r_2^L\, , \, p_0 r_3^H + (1-p_0) r_3^L$	$p_1 r_1^H + (1-p_1) r_1^L\, , \, p_1 r_2^H + (1-p_1) r_2^L - c\, , \, p_1 r_3^H + (1-p_1) r_3^L$
	1	$p_1 r_1^H + (1-p_1) r_1^L - c\, , \, p_1 r_2^H + (1-p_1) r_2^L\, , \, p_1 r_3^H + (1-p_1) r_3^L$	$p_2 r_1^H + (1-p_2) r_1^L - c\, , \, p_2 r_2^H + (1-p_2) r_2^L - c\, , \, p_2 r_3^H + (1-p_2) r_3^L$

Agent 3: $e_3 = 0$

		Agent 2	
		0	1
Agent 1	0	$p_1 r_1^H + (1-p_1) r_1^L\, , \, p_1 r_2^H + (1-p_1) r_2^L\, , \, p_1 r_3^H + (1-p_1) r_3^L - c$	$p_2 r_1^H + (1-p_2) r_1^L\, , \, p_2 r_2^H + (1-p_2) r_2^L - c\, , \, p_2 r_3^H + (1-p_2) r_3^L - c$
	1	$p_2 r_1^H + (1-p_2) r_1^L - c\, , \, p_2 r_2^H + (1-p_2) r_2^L\, , \, p_2 r_3^H + (1-p_2) r_3^L - c$	$p_2 r_1^H + (1-p_2) r_1^L - c\, , \, p_2 r_2^H + (1-p_2) r_2^L - c\, , \, p_2 r_3^H + (1-p_2) r_3^L - c$

Agent 3: $e_3 = 1$

Figure 3.2: Three-agent game G_2

PROPOSITION 11 (Optimal three-agent mechanism: Unique Nash implementation). *An optimal reward scheme inducing two agents (agents 1 and 2) to exert effort and the third agent to shirk is as follows:*

$$
\begin{aligned}
&W^U = \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1} \\
&r_1^H = \frac{c}{p_1 - p_0}\,, \quad r_2^H = \frac{c}{p_2 - p_1}\,, \quad r_3^H = 0 \\
&r_1^L = 0\,, \quad r_2^L = 0\,, \quad r_3^L = \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1}.
\end{aligned}
$$

In this reward scheme, agents 1 and 2 are treated asymmetrically, while agent 3 picks up the entire budget when the team performs poorly.

Use of an alternative beneficiary for diversion of resources, in case the originally intended group does not perform well, is not that uncommon in organizations: annual bonus allocated for one business unit may be shifted to another unit with the former failing to meet a threshold performance level.

One drawback of the above implementation mechanism is that agents 1 and 2 are treated asymmetrically. For a principal concerned about fairness issues, there are two ways to deal with this problem to which we turn next.

■ *Weak Nash implementation.* In the unique Nash implementation mechanism above, agent 1 receives a higher reward than agent 2 so that exerting effort is a dominant strategy for agent 1 and in response agent 2 exerts effort. It is possible, however, to lower agent 1's reward down to agent 2's reward and induce him to exert effort so long as agent 2 is also exerting effort. This might open the door for a second Nash equilibrium with both agents shirking. Thus while implementation costs might be lower with symmetry between the two players restored, the uniqueness of equilibrium can no longer be guaranteed. Formally, for weak Nash implementation of $(1, 1, 0)$ the incentive compatibility conditions are less stringent for agent 1 as follows:

$$
\begin{aligned}
&[\mathbf{IC}_1'] \qquad r_1^H - r_1^L \geq \frac{c}{p_2 - p_1}, \\
&[\mathbf{IC}_2'] \qquad r_2^H - r_2^L \geq \frac{c}{p_2 - p_1}.
\end{aligned}
$$

Using the [IC′] conditions, and the same participation constraints ([PC$_1$] and [PC$_2$]) and non-negative conditions on the rewards as in the problem $(\mathcal{P})$,[8] the optimal mechanism can be characterized as follows (derivations can be found in the Appendix):

$$
\begin{aligned}
&W^N = \frac{2c}{p_2 - p_1} \\
&r_1^H = r_2^H = \frac{c}{p_2 - p_1}\,, \quad r_3^H = 0 \\
&r_1^L = r_2^L = 0\,, \quad r_3^L = \frac{2c}{p_2 - p_1}.
\end{aligned}
$$

It can be checked that for the above incentives, both $(1,1,0)$ and $(0,0,0)$ are Nash equilibria. Thus the principal gains in terms of fairness and cost efficiency but loses control due to the coordination problem.

■ *Unique symmetric implementation.* One may be interested in a *symmetric mechanism* treating only agents 1 and 2 identically that implements the desired effort profile uniquely.[9] For uniqueness, the incentive compatibility conditions for the agents should be the same as [IC$_1$] and [IC$_2$] in the problem $(\mathcal{P})$. For a *weak* form of symmetry, we require that the rewards to agents 1 and 2 be equal for high signal realization while for low signal the rewards can be different. Then the principal's cost minimization yields the following result (derivations can be found in the

[8] It can be easily verified that agent 3's IC and PC constraints will be satisfied.

[9] Later on in section 5 we extend symmetric treatment of rewards to agent 3 as well, upon high signal, that we call *complete symmetry*.

Appendix):

$$\begin{aligned}
&W^S = \frac{2c}{p_1 - p_0} \\
&r_1^H = \frac{c}{p_1 - p_0}\,, \quad r_2^H = \frac{c}{p_1 - p_0}\,, \quad r_3^H = 0 \\
&r_1^L = 0\,, \quad r_2^L \in \left[0,\, \frac{c}{p_1 - p_0} - \frac{c}{p_2 - p_1}\right]\,, \quad r_3^L = \frac{2c}{p_1 - p_0} - r_2^L.
\end{aligned}$$

From the above solution, one can see that the combined requirement of uniqueness of Nash equilibrium and (weakly) symmetric treatment of agents 1 and 2 lead to principal's implementation costs being the highest among the three mechanisms studied here. Also note that if r_2^L is set at 0 (along with $r_1^L = 0$), exerting effort is a *dominant strategy* for both agents 1 and 2.

In all (weakly) symmetric mechanisms, the incentive compatibility condition need not be binding for agent 2. The observation is true because, symmetric rewards upon high signal is an additional requirement which distinguishes the problem from unique Nash implementation. Therefore, from the optimal mechanism in Proposition 11, r_2^H is pushed up to match agent 1's reward, which creates a room for flexibility for r_2^L. In this sense, incentive compatibility constraint may be slack and thus some surplus may accrue to agent 2. From the perspective of the principal, pursuing fairness is costly and makes the mechanism suboptimal among the class of unique implementation mechanisms.

Table 3.1 summarizes comparison of the different mechanisms analyzed.

Implementation	Cost	Uniqueness	Symmetry between 1 & 2
Unique Nash	Medium	✓	✗
Weak Nash	Low	✗	✓
Unique Symmetric	High	✓	✓ (weak)
Dominant Strategies	High	✓	✓

Table 3.1: Trade-offs among three-agent mechanisms

3.4 Money burning or sabotage: Choosing between two evils

So far the third agent has acted merely to endogenize money burning without actually burning money. This idea, while it originally appeared in MacLeod (2003) in a principal-agent setting, we integrate explicitly in the teams problem. Our main objective is to look beyond just the facilitating role of agent 3 in endogenizing money burning. Given that agent 3 stands to benefit when the team performs poorly suggests that it should be in his interest to *sabotage* team performance. So, in this section, we model sabotage formally and look at its implications. We will see that expanding the role of agent 3 to engage in sabotage may expose the limitations of subjective evaluations and money burning in a way that has not been considered in MacLeod (2003), or in the related relations contracting model of Levin (2003).

Let us start with the modification that agent 3 can now exert one unit of *sabotage effort*, which costs the same as the productive effort, to reduce the team's chances of sending a high performance signal from p_0, p_1, p_2 to some respectively lower values p'_0, p'_1, p'_2 satisfying $p'_0 < p'_1 < p'_2$.[10] We do not consider the possibility of sabotage by agents 1 or 2

[10] So long as the cost of sabotage effort does not exceed the cost of productive effort, our main negative result in this section (Proposition 12) will be unaffected. Obviously if sabotage effort is more costly, its attractiveness for the saboteur will lessen.

as not only they have nothing to gain but will definitely lose from such activities. We call this game, the *sabotage-augmented game*.

Our first result is a negative one casting doubts on the implementability of full efforts objective:

LEMMA 5. *Given any three-agent mechanism, full team efforts,* $(1,1,0)$*, is not implementable due to agent 3's deviation to sabotage if and only if* [11]

$$(p_2 - p_2')(r_3^L - r_3^H) \,>\, c \tag{3.4}$$

holds.

Proof. Given a reward scheme, the LHS of (3.4) is the benefit of sabotage for agent 3 while the RHS denotes the cost, and whenever benefit exceeds cost the agent will deviate and sabotage. This can also be directly verified by rewriting (3.4) as

$$p_2 r_3^H + (1-p_2) r_3^L \,<\, p_2' r_3^H + (1-p_2') r_3^L - c,$$

which implies the agent gets more reward from sabotage than by simply shirking. □

To see that some of the implementation mechanisms derived in the earlier section might fail due to sabotage, consider the *unique Nash implementation mechanism*:

$$r_3^L - r_3^H = \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1}.$$

Now if (and only if)

$$\frac{c}{p_2 - p_2'} \,<\, \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1} \tag{3.5}$$

[11] When agent 3 is indifferent between sabotage and not sabotage, which would then imply he exerts zero effort, we assume that he will break the indifference by not sabotaging.

holds, then it is clear that the three-agent mechanism that was constructed without any consideration of sabotage will indeed open the door to sabotage by agent 3. Note that this last condition is purely a technological condition, after dropping the variable c. The condition is likely to hold if the drop in the probability of high signal following sabotage is significant so that $p_2 - p_2'$ is sufficiently high. Similar conditions can be obtained for sabotage for the other mechanisms derived in section 3: the contributory role of $p_2 - p_2'$ remains the same, only the incremental reward due to sabotage, $r_3^L - r_3^H$, will differ.

Generally, if condition (3.4) holds, we cannot implement full efforts $(1, 1, 0)$ unless the rewards are modified to construct a sabotage-proof mechanism. With that objective in mind, denote $e_3 = -1$ as one unit of sabotage effort. The new normal form game is shown in Fig. 3.3. Lemma 4 still applies and exerting one unit of productive effort (working) is a dominated strategy for agent 3. So we only need to consider whether agent 3 will shirk or sabotage. If for *some* incentive mechanism the strategy profile $(e_1, e_2, e_3) = (1, 1, 0)$ can be implemented, we say that the mechanism is *sabotage-proof.*

Next we are going to argue that given condition (3.4) holds (i.e., given that the originally derived mechanisms are *not* sabotage-proof), a sabotage-proof mechanism does not exist in any of the implementation cases analyzed in section 3.3, except for the unique symmetric implementation.

■ *Non-existence of a sabotage-proof mechanism.* Note that the occurrence of sabotage for any given mechanism depends on its relative cost and benefit specific to the mechanism. For example, if (3.5) does not hold, the *unique Nash implementation mechanism* is essentially a sabotage-proof mechanism. But what we are interested in is the complementary situation of condition (3.5), i.e., allowing that sabotage *does*

		Agent 2	
		0	1
Agent 1	0	$p'_0 r_1^H + (1-p'_0) r_1^L\,,\ p'_0 r_2^H + (1-p'_0) r_2^L\,,\ p'_0 r_3^H + (1-p'_0) r_3^L - c$	$p'_1 r_1^H + (1-p'_1) r_1^L\,,\ p'_1 r_2^H + (1-p'_1) r_2^L - c\,,\ p'_1 r_3^H + (1-p'_1) r_3^L - c$
	1	$p'_1 r_1^H + (1-p'_1) r_1^L - c\,,\ p'_1 r_2^H + (1-p'_1) r_2^L\,,\ p'_1 r_3^H + (1-p'_1) r_3^L - c$	$p'_2 r_1^H + (1-p'_2) r_1^L - c\,,\ p'_2 r_2^H + (1-p'_2) r_2^L - c\,,\ p'_2 r_3^H + (1-p'_2) r_3^L - c$

Agent 3: $e_3 = -1$

		Agent 2	
		0	1
Agent 1	0	$p_0 r_1^H + (1-p_0) r_1^L\,,\ p_0 r_2^H + (1-p_0) r_2^L\,,\ p_0 r_3^H + (1-p_0) r_3^L$	$p_1 r_1^H + (1-p_1) r_1^L\,,\ p_1 r_2^H + (1-p_1) r_2^L - c\,,\ p_1 r_3^H + (1-p_1) r_3^L$
	1	$p_1 r_1^H + (1-p_1) r_1^L - c\,,\ p_1 r_2^H + (1-p_1) r_2^L\,,\ p_1 r_3^H + (1-p_1) r_3^L$	$p_2 r_1^H + (1-p_2) r_1^L - c\,,\ p_2 r_2^H + (1-p_2) r_2^L - c\,,\ p_2 r_3^H + (1-p_2) r_3^L$

Agent 3: $e_3 = 0$

		Agent 2	
		0	1
Agent 1	0	$p_1 r_1^H + (1-p_1) r_1^L\,,\ p_1 r_2^H + (1-p_1) r_2^L\,,\ p_1 r_3^H + (1-p_1) r_3^L - c$	$p_2 r_1^H + (1-p_2) r_1^L\,,\ p_2 r_2^H + (1-p_2) r_2^L - c\,,\ p_2 r_3^H + (1-p_2) r_3^L - c$
	1	$p_2 r_1^H + (1-p_2) r_1^L - c\,,\ p_2 r_2^H + (1-p_2) r_2^L\,,\ p_2 r_3^H + (1-p_2) r_3^L - c$	$p_2 r_1^H + (1-p_2) r_1^L - c\,,\ p_2 r_2^H + (1-p_2) r_2^L - c\,,\ p_2 r_3^H + (1-p_2) r_3^L - c$

Agent 3: $e_3 = 1$

Figure 3.3: Three-agent game with sabotage

destroy the implementability of a derived mechanism, whether there is any way to remedy the situation. Let us now assume the optimistic scenario that there indeed exists an alternative sabotage-proof mechanism, and use $\hat{r}_i$ to denote the reward scheme specified by this mechanism. From Lemma 5 it follows that agent 3 will be deterred from sabotage if only if

$$(p_2 - p_2')(\hat{r}_3^L - \hat{r}_3^H) \leq c. \tag{3.6}$$

Given that agent 3 will exert zero effort rather than sabotage in the posited sabotage-proof equilibrium, agents 1 and 2's incentive compatibility conditions are similar to ones in the case without sabotage. Let us consider once again the *unique Nash implementation* mechanism of section 3 for illustration:

$$\begin{aligned} &\textbf{[IC}_1\textbf{]} \qquad \hat{r}_1^H - \hat{r}_1^L \geq \frac{c}{p_1 - p_0}, \\ &\textbf{[IC}_2\textbf{]} \qquad \hat{r}_2^H - \hat{r}_2^L \geq \frac{c}{p_2 - p_1}. \end{aligned}$$

However, given the budget constraint as in (3.3), we have the following:

$$\left[\hat{r}_1^H - \hat{r}_1^L\right] + \left[\hat{r}_2^H - \hat{r}_2^L\right] = \hat{r}_3^L - \hat{r}_3^H. \tag{3.7}$$

Therefore,

$$\begin{aligned} \frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1} &\leq \left[\hat{r}_1^H - \hat{r}_1^L\right] + \left[\hat{r}_2^H - \hat{r}_2^L\right] \\ &= \hat{r}_3^L - \hat{r}_3^H \\ &\leq \frac{c}{p_2 - p_2'}, \qquad \text{(by (3.6))} \end{aligned}$$

i.e.,

$$\frac{c}{p_1 - p_0} + \frac{c}{p_2 - p_1} \leq \frac{c}{p_2 - p_2'}. \tag{3.8}$$

This implies that a sabotage-proof mechanism exists only if (3.8) is true.

But then it contradicts with (3.5), which is our starting hypothesis that sabotage will occur in the originally derived unique Nash implementation mechanism of section 3. Therefore, once sabotage occurs, we are no longer able to rectify it by any alternative reward scheme. A similar analysis for weak Nash and dominant strategy implementations will lead to the same conclusion. The following proposition summarizes one of our main results.

PROPOSITION 12 (Non-existence of Sabotage-proof mechanism). *For (weak or unique) Nash or dominant strategy implementation, if the optimal mechanism derived in section 3 where sabotage was not considered fails to implement efforts from agents 1 and 2 and deter sabotage by agent 3 in the sabotage-augmented game, then there does not exist any sabotage-proof mechanism that can rectify the problem of sabotage.*

The intuition for Proposition 12 is as follows. In the game without any consideration of sabotage, the principal could freely transfer the rewards from agents 1 and 2 to agent 3 upon bad performance in order to incentivize agents 1 and 2 to exert efforts without worrying about agent 3 undoing the incentives. This allowed the principal to tighten the incentive compatibility conditions of agents 1 and 2 to bind in the optimal mechanism derived in section 3. However, once sabotage is allowed, agent 3's reward has to be restricted too such that his benefit from sabotage is (weakly) below his sabotage effort cost (see (3.6)). But this goes against agents 1 and 2's incentives: LHS of (3.7) determines, collectively, effort incentives of agents 1 and 2 that equals the sabotage incentive of agent 3 (RHS). The higher is the potential benefit of sabotage to agent 3 ($p_2 - p_2'$ is large), the more restrictive are the incentives for agents 1 and 2, which finally breaks down the incentive compatibility conditions for inducing efforts. The incentives provided by the original optimal mechanisms already reached the lower bound of the combined IC conditions for agents 1 and 2 that had induced agent 3 to sabotage. Then in order to deter sabotage by decreasing agent 3's incentive, which equals the sum of agents 1 and 2's incentives, any attempt will lead to insufficient incentives for one or both of them.

■ *A special case of deterring sabotage.* Proposition 12 shows the tension between incentivizing agents 1 and 2 for efforts and dissuading agent 3 against sabotage. This tension was finely balanced such that agents 1 and 2 were indifferent between exerting effort and shirking. However, in the case of *unique symmetric implementation*, for the optimal mechanism in section 3 agent 2 *strictly prefers* to exert effort over shirking while agent 1 is indifferent. This means there is still some room to transfer rewards back from agent 3 to agent 2 following a low signal. Now, if sabotage occurs in the optimal mechanism, with a marginal transfer from agent 3 back to agent 2 the latter's effort incentive can be maintained at least for a while. This gives rise to the possibility of restoring sabotage-proofness. In the following we show this possibility.

Consider the following implementation mechanism as derived in section 3 with $r_2^L = \Delta$, where $0 < \Delta < \frac{c}{p_1-p_0} - \frac{c}{p_2-p_1}$:

$$\text{[Mechanism M°]} \qquad \begin{cases} W^S = \dfrac{2c}{p_1 - p_0} \\ r_1^H = \dfrac{c}{p_1 - p_0}, \quad r_2^H = \dfrac{c}{p_1 - p_0}, \quad r_3^H = 0 \\ r_1^L = 0, \quad r_2^L = \Delta, \quad r_3^L = \dfrac{2c}{p_1 - p_0} - \Delta. \end{cases}$$

Further suppose sabotage arises given M° so that condition (3.4) can be rewritten as:

$$\frac{c}{p_2 - p_2'} < \frac{2c}{p_1 - p_0} - \Delta. \tag{3.9}$$

Now we claim that for some technologies, i.e., some values of $\frac{c}{p_2 - p_2'}$, the rewards can be adjusted to induce efforts by agents 1 and 2 and deter sabotage by agent 3, even without pushing up the principal's cost. Let

$$\frac{c}{p_2 - p_2'} = \frac{3}{2}\frac{c}{p_1 - p_0} + \frac{1}{2}\frac{c}{p_2 - p_1} - \frac{\Delta}{2},$$

which satisfies condition (3.9). Then consider the following mechanism:

$$\text{[Mechanism } \hat{\mathrm{M}}\text{]} \qquad \begin{cases} \hat{W}^S = \dfrac{2c}{p_1 - p_0} \\ \hat{r}_1^H = \dfrac{c}{p_1 - p_0}, \quad \hat{r}_2^H = \dfrac{c}{p_1 - p_0}, \quad \hat{r}_3^H = 0 \\ \hat{r}_1^L = 0, \quad \hat{r}_2^L = \dfrac{c}{p_1 - p_0} - \dfrac{c}{p_2 - p_1}, \quad \hat{r}_3^L = \dfrac{c}{p_1 - p_0} + \dfrac{c}{p_2 - p_1}. \end{cases}$$

It can be verified that [IC$_1$], [IC$_2$] and agent 3's no-sabotage condition (3.6) are all satisfied, while the principal's cost remains the same as in M°. Thus, sabotage is deterred under the new mechanism $\hat{\mathrm{M}}$.[12] Another way to deter sabotage in this example would be to reduce r_2^H rather than increase r_2^L. This may also lead to a sabotage-proof mechanism but it breaks the symmetric treatment of agents 1 and 2.

■ *Collusion between agent 3 and any of the other agents.* That potential collusion between the parties who are being incentivized could pose a problem has earlier been noted by Esawaran and Kotwal (1984) in a standard verifiable production team problem. Besides direct sabotage, which is the main focus in this paper, collusion can pose a similar problem in our setting. To see how, consider a side-deal between agent 3 and any of agents 1 and 2 who are expected to put in efforts in the optimal incentives. If agent 3 can convince them that following poor performance they will be more than adequately compensated for the loss in rewards by agent 3, then they will shirk. This will increase the possibility of poor team performance and if that eventuality were to happen, all of them will have saved up their effort costs – agent 3 not engaging in sabotage and

[12] Note that $\hat{\mathrm{M}}$ itself is an optimal unique symmetric implementation mechanism without considering sabotage, which is shown to be sabotage-proof given that M° is not.

the other agents not engaging in team production. This will upset the principal's objective of saving money burning through budget balance and favoritism and totally undermine the effectiveness of subjective performance evaluation. This problem is different from team performance being low in the verifiable team production model of Eswaran and Kotwal (1984), as for the latter poor team performance means there will be very little to share among the team members, whereas in our case principal-committed rewards being fixed all three agents stand to gain collectively at the principal's expense. Thus, collusion poses a greater challenge in the subjective evaluation formulation of the team problem than in the standard verifiable production team problem. The principal will now face the challenge of double incentivization: deter sabotage as well as collusion.[13] But since we already pointed out that sabotage-proofness might fail, achieving both collusion-proofness and sabotage-proofness becomes a difficult proposition.

3.5 Variations of three-agent model without sabotage

In this section, we study some variations of the three-agent model to endogenize money burning (but without the considerations of sabotage) to approximate more closely real-life organizations.

■ *Favoritism*. In the three-agent mechanism, the third agent is favored by the principal on receipt of a bad performance signal. However, in many applications there might be a clear favorite of the boss, something if unrelated to superior productivity is viewed as nepotism. Since in our model one cannot ascribe team performance one-to-one to any specific

[13] Deterring collusion means the principal will have to undertake costly monitoring to make side-contracting difficult.

agent's efforts, we will analyze the case of clear favoritism to be one where one agent, say agent 3, receives a higher reward than any other agent under all circumstances:

$$r_3^H \geq \max\{r_1^H, r_2^H\} \quad \text{and} \quad r_3^L \geq \max\{r_1^L, r_2^L\}.$$

The incentive compatibility and participation constraints remain the same as in the problem $(\mathcal{P})$ (agent 3's IC and PC can also be easily verified). Solving the principal's problem leads to:

$$\begin{aligned}
&W^{FM} = \frac{2c}{p_1 - p_0} + \frac{c}{p_2 - p_1} \\
&r_1^H = \frac{c}{p_1 - p_0}, \quad r_2^H = \frac{c}{p_2 - p_1}, \quad r_3^H = \frac{c}{p_1 - p_0} \\
&r_1^L = 0, \quad r_2^L = 0, \quad r_3^L = \frac{2c}{p_1 - p_0} + \frac{c}{p_2 - p_1}.
\end{aligned}$$

Principal ranks the agents according to: Agent 3 $\succeq$ Agent 1 $\succeq$ Agent 2. The marginal reward upon good performance is the highest for agent 1 who exerts effort in a dominant strategy, followed by a lower marginal reward for agent 2 who exerts effort as a Nash best response to agent 1's effort and the lowest marginal reward is for the most favored agent 3 whose dominant strategy is to shirk. Here agent 3 can either exert effort positively or shirk but cannot sabotage. Compared with the unique Nash implementation mechanism of Proposition 11, the principal spends more resources due to the additional payment to the favored agent.

■ *Symmetry.* As another extreme, consider a completely symmetric treatment among the agents regardless of high or low performance signal: $r_i^H = r_i^L = \frac{W}{3}$. In this case, none will have any incentive to work since effort only incurs a cost to an agent but never generates additional rewards. A more sensible situation is to treat the agents equally if performance is good but reward them differently upon bad performance.

Again, let agent 3 be the one who is not expected to work. Then symmetric treatment implies that the reward for agents 1, 2 and 3 should be the same if a high performance signal is reported:

$$r_1^H = r_2^H = r_3^H .$$

If a low performance signal is reported, the favoritism condition still applies:

$$r_3^L \geq \max\{r_1^L , r_2^L\}.$$

Solving the principal's cost minimization problem subject to these constraints as well as the same [ICs] and [PCs] conditions yields a stronger form of *complete symmetric* mechanism:

$$\begin{aligned}
&W^{SC} = \frac{3c}{p_1 - p_0} \\
&r_1^H = r_2^H = r_3^H = \frac{c}{p_1 - p_0} \\
&r_1^L = 0\,,\ r_2^L \in \left[0, \frac{c}{p_1 - p_0} - \frac{c}{p_2 - p_1}\right]\,,\ r_3^L = \frac{3c}{p_1 - p_0} - r_2^L .
\end{aligned}$$

This mechanism imposes strict symmetry across all agents upon high performance signal, which is more fair and realistic than the previously derived favoritism mechanism. However, it costs the principal highest among all mechanisms we have studies so far.

The main message to take away is that as the principal introduces more and more symmetry in the treatment among team members, implementing desired efforts becomes costly. This is to be expected given that maintaining budget balance (and thus avoiding money burning) together with inducement of agent efforts is impossible if one starts from full symmetry. Thus the tradeoff between fairness and cost efficiency remains a key problem for the principal under subjective performance evaluation, just like the tension pointed out in Winter's (2004) article even when team performance could be verified.

3.6 Conclusion

Our analysis in this paper conveys a negative message for incentives in organizations, where a group of agents need to work cooperatively to a common team goal but team performance not measurable. Due to complementarity between agents' efforts, economic efficiency dictates that identical agents be treated asymmetrically. In addition, subjectivity in performance assessment implies organizations must be prepared to throw away resources to incentivize the agents. This calls for further discriminatory treatment and even blatant favoritism, if wastage of resources are to be avoided. But even if one were to sacrifice fairness for economic efficiency, the loss in economic efficiency may be unavoidable due to perverse sabotage incentives generated due to discrimination.

One difference between the concern for budget balance in team production with objective performance measure and budget balance under subjective evaluation as in this paper is that, in the former total *team output* must be distributed exhaustively among the team members whereas in this paper *principal-committed rewards* must be distributed. Both pose problems for the organization: in the first case the main concern is implementation of first-best efforts, whereas in ours the difficulty is to induce all important team members to exert efforts. Ensuring budget balance with the help of a third party tends to undermine both these objectives. Ultimately, the agents may have to be monitored for any subversive activity that could undermine the team's objective.

APPENDIX A

Proofs for Chapter 1

Proof of Proposition 2. The agent's payoffs in the repeated efforts game are as follows:

$$
\begin{aligned}
V(1,1) &= W - p_1p_1z^{HH} - p_1(1-p_1)z^{HL} - (1-p_1)p_1z^{LH} - (1-p_1)(1-p_1)z^{LL} - 2c \\
V(1,0) &= W - p_1p_0z^{HH} - p_1(1-p_0)z^{HL} - (1-p_0)p_1z^{LH} - (1-p_1)(1-p_0)z^{LL} - c \\
V(0,1) &= W - p_0p_1z^{HH} - p_0(1-p_1)z^{HL} - (1-p_1)p_0z^{LH} - (1-p_0)(1-p_1)z^{LL} - c \\
V(0,0) &= W - p_0p_0z^{HH} - p_0(1-p_0)z^{HL} - (1-p_0)p_0z^{LH} - (1-p_0)(1-p_0)z^{LL}.
\end{aligned}
$$

For full efforts implementation, the agent's incentive compatible conditions are:

$$p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) \geq \frac{c}{p_1 - p_0} \tag{A.1}$$

$$p_1(z^{LH} - z^{HH}) + (1-p_1)(z^{LL} - z^{HL}) \geq \frac{c}{p_1 - p_0} \tag{A.2}$$

$$p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) + p_0(z^{LH} - z^{HH}) + (1-p_0)(z^{LL} - z^{HL}) \geq \frac{2c}{p_1 - p_0}. \tag{A.3}$$

Thus the principal's problem can be written as follows:

$$\min_{\{W, z^{HH}, z^{HL}, z^{LH}, z^{LL}\}} \quad W$$

$$\begin{aligned} s.t. \qquad & \text{(A.1)}, \text{(A.2)}, \text{(A.3)} \\ & z^{HH}, z^{HL}, z^{LH}, z^{LL} \geq 0 \\ & W - z^{HH}, W - z^{HL}, W - z^{LH}, W - z^{LL} \geq 0. \end{aligned}$$

Write the Lagrangian:

$$\begin{aligned} L = & -W + A\left[p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) - \frac{c}{p_1 - p_0}\right] \\ & + B\left[p_1(z^{LH} - z^{HH}) + (1-p_1)(z^{LL} - z^{HL}) - \frac{c}{p_1 - p_0}\right] \\ & + C\left[p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) + p_0(z^{LH} - z^{HH}) + (1-p_0)(z^{LL} - z^{HL}) - \frac{2c}{p_1 - p_0}\right] \\ & + Dz^{HH} + Ez^{HL} + Fz^{LH} + Gz^{LL} \\ & + H(W - z^{HH}) + I(W - z^{HL}) + J(W - z^{LH}) + K(W - z^{LL}). \end{aligned}$$

First-order conditions are:

$$\frac{\partial L}{\partial W} = -1 + H + I + J + K = 0 \tag{A.4a}$$

$$\frac{\partial L}{\partial z^{HH}} = -Ap_1 - Bp_1 - Cp_1 - Cp_0 + D - H = 0 \tag{A.4b}$$

$$\frac{\partial L}{\partial z^{HL}} = Ap_1 - B(1-p_1) + Cp_1 - C(1-p_0) + E - I = 0 \tag{A.4c}$$

$$\frac{\partial L}{\partial z^{LH}} = -A(1-p_1) + Bp_1 - C(1-p_1) + Cp_0 + F - J = 0 \tag{A.4d}$$

$$\frac{\partial L}{\partial z^{LL}} = A(1-p_1) + B(1-p_1) + C(1-p_1) + C(1-p_0) + G - K = 0. \tag{A.4e}$$

Since $D \geq 0$, suppose $D = 0$. From (A.4b), $D = Ap_1 + Bp_1 + C(p_1 + p_0) + H$, which implies $A = B = C = H = 0$. Then we have the following:

$$\begin{cases} I + J + K = 1 & \text{by (A.4a)} \\ E = I\,,\ F = J\,,\ G = K & \text{by (A.4c), (A.4d), (A.4e).} \end{cases}$$

Since $I, J, K \geq 0$, at least one of I, J or K should be strictly positive. For the time being, suppose $I > 0$. Then we have $E = I > 0$, thus

$$z^{HL} = 0 \text{ and } W - z^{HL} = 0,$$

which implies $W = z^{HH} = z^{HL} = z^{LH} = z^{LL} = 0$, contradicting (A.1), (A.2) and (A.3). So $E = I = 0$. If alternatively we suppose J or $K > 0$, the same argument applies. Therefore, $D = 0$ should not hold and we must have $D > 0$, which implies

$$z^{HH} = 0.$$

Then we have $W - z^{HH} > 0$, which implies $H = 0$.

Now (A.4b) can be written as $D = Ap_1 + Bp_1 + C(p_1 + p_0) > 0$, and at least one of A, B or C should be strictly positive. Then by (A.4e), $K = A(1-p_1) + B(1-p_1) + C(1-p_1) + C(1-p_0) + G > 0$. This implies $W - z^{LL} = 0$, thus

$$W = z^{LL} > 0 \text{ and } G = 0.$$

So far the first-order conditions can be simplified as follows with $D > 0$ and $K > 0$:

$$\textbf{[FOCs]} \quad \begin{cases} D = Ap_1 + Bp_1 + C(p_1 + p_0) & D + E = B + C + I \\ I + J + K = 1 & D + F = A + C + J \\ D + E + F = 1 & D + K = A + B + 2C. \end{cases}$$

With at least one of A, B or C being strictly positive in mind, we discuss different cases in the next steps.

i. $A > 0\,,\ B = C = 0$. **[FOCs]** become

$$\begin{cases} D = Ap_1 > 0 \\ D + E = I\ >\ 0 \quad \Rightarrow \quad W - z^{HL} = 0 \quad \Rightarrow \quad W = z^{HL} = z^{LL} \\ D + F = A + J \quad \Rightarrow \quad F = A(1-p_1) + J\ >\ 0 \quad \Rightarrow \quad z^{LH} = 0. \end{cases}$$

Using these values in (A.2) leads to a contradiction, so there is no solution in this case.

ii. $B > 0\,,\ A = C = 0$. **[FOCs]** become

$$\begin{cases} D = Bp_1 > 0 \\ D + E = B + I \quad \Rightarrow \quad E = B(1-p_1) + I\ >\ 0 \quad \Rightarrow \quad z^{HL} = 0 \\ D + F = J\ >\ 0 \quad \Rightarrow \quad W - z^{LH} > 0 \quad \Rightarrow \quad W = z^{LH} = z^{LL}. \end{cases}$$

Using these values in (A.1) leads to a contradiction, so there is no solution in this case.

iii. $C > 0\,,\ A = B = 0$. This implies $D = C(p_1 + p_0)$, so we have the following subcases.

(a) $p_1 + p_0 = 1$. This implies $D = C\,,\ E = I\,,\ F = J$. Suppose $E = I > 0$, and this leads to the same contradiction as discussed earlier. Therefore, we must have $E = I = 0$, and similarly $F = J = 0$. Then we can further conclude that $C = D = K = 1$. Since (A.3) is binding (by $C = 1 > 0$), we have:

$$p_1(z^{HL}-z^{HH})+(1-p_1)(z^{LL}-z^{LH})+p_0(z^{LH}-z^{HH})+(1-p_0)(z^{LL}-z^{HL}) = \frac{2c}{p_1 - p_0},$$

which implies

$$W = z^{LL} = \frac{2c}{p_1 - p_0}. \qquad \text{(by } p_1 + p_0 = 1)$$

Then consider (A.1) and (A.2). By $A = B = 0$ and $p_1 + p_0 = 1$, we have the following:

$$\begin{cases} p_1(z^{HL} - z^{HH}) + (1 - p_1)(z^{LL} - z^{LH}) \geq \frac{c}{p_1 - p_0} & \Rightarrow \quad p_1 z^{HL} \geq p_0 z^{LH} + c \\ p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) \geq \frac{c}{p_1 - p_0} & \Rightarrow \quad p_0 z^{HL} \leq p_1 z^{LH} - c, \end{cases}$$

which imply

$$z^{HL} \geq \frac{c}{p_1 - p_0} \quad \text{and} \quad z^{LH} \geq \frac{c}{p_1 - p_0}.$$

This yields the following **solution** for the case of $p_1 + p_0 = 1$:

$$\begin{cases} W = \dfrac{2c}{p_1 - p_0} \\ z^{LH}, z^{HL} \in [\dfrac{c}{p_1 - p_0}, \dfrac{2c}{p_1 - p_0}] \text{ and } p_1 z^{HL} \geq p_0 z^{LH} + c\,,\, p_0 z^{HL} \geq p_1 z^{LH} - c \\ z^{HH} = 0\,,\ z^{HL} = z^{LH} = \dfrac{c}{p_1 - p_0}\,,\ z^{LL} = \dfrac{2c}{p_1 - p_0}. \end{cases}$$

(b) $p_1 + p_0 > 1$. This implies

$$\begin{cases} C(p_1 + p_0 - 1) + E = I \ > \ 0 \ \Rightarrow W = z^{HL} \ > \ 0 \ \Rightarrow \ E = 0 \\ C(p_1 + p_0 - 1) + F = J \ > \ 0 \ \Rightarrow W = z^{LH} \ > \ 0 \ \Rightarrow \ F = 0. \end{cases}$$

Since (A.3) is binding (by $C > 0$),

$$p_1(z^{HL} - z^{HH}) + (1 - p_1)(z^{LL} - z^{LH}) + p_0(z^{LH} - z^{HH}) + (1 - p_0)(z^{LL} - z^{HL}) = \frac{2c}{p_1 - p_0}.$$

Using $W = z^{HL} = z^{LH} = z^{LL}$ and $z^{HH} = 0$ in the above equation, we have:

$$W = z^{HL} = z^{LH} = z^{LL} = \frac{2c}{(p_1 + p_0)(p_1 - p_0)}.$$

It can be verified that these values will satisfy the constraints

(A.1) and (A.2). We thus obtain the **solution** for the case of $p_1 + p_0 > 1$:

$$\begin{cases} W = \dfrac{2c}{(p_1+p_0)(p_1-p_0)} \\ z^{HH} = 0\,,\ z^{HL} = z^{LH} = z^{LL} = \dfrac{2c}{(p_1+p_0)(p_1-p_0)}. \end{cases}$$

(c) $p_1 + p_0 < 1$. This implies

$$\begin{cases} E = C(1-p_1-p_0) + I \ > \ 0 \ \Rightarrow z^{HL} \ = \ 0 \ \Rightarrow W - z^{HL} > 0 \Rightarrow \ I = 0 \\ F = C(1-p_1-p_0) + J \ > \ 0 \ \Rightarrow z^{LH} \ = \ 0 \ \Rightarrow W - z^{LH} > 0 \Rightarrow \ J = 0. \end{cases}$$

Using $z^{HH} = z^{HL} = z^{LH} = 0$ and $W = z^{LL}$ in the IC condition (A.3) with equality (by $C > 0$), we have:

$$W = z^{LL} = \frac{2c}{(2-p_1-p_0)(p_1-p_0)}.$$

Using these values, the IC conditions (A.1) and (A.2) will be violated, a contradiction. Hence, there is no solution in this case.

iv. $A > 0\,,\ C > 0\,,\ B = 0$. This implies $D = Ap_1 + C(p_1+p_0)$, and we need to discuss different cases for $p_1 + p_0$ again.

(a) $p_1 + p_0 = 1$. This implies

$$\begin{cases} I = Ap_1 + E \ > \ 0 \ \Rightarrow W = z^{HL} \\ F = A(1-p_1) + J \ > \ 0 \ \Rightarrow z^{LH} \ = \ 0 \ \Rightarrow W - z^{LH} > 0 \Rightarrow \ J = 0. \end{cases}$$

Using $z^{LH} = z^{HH} = 0$ and $W = z^{HL} = z^{LL}$ in (A.2) leads to a contradiction, so there is no solution in this case.

(b) $p_1 + p_0 > 1$. This implies

$$I = Ap_1 + C(p_1+p_0-1) + E \ > \ 0 \ \Rightarrow W = z^{HL} \ > \ 0 \ \Rightarrow \ E = 0.$$

Using $W = z^{HL} = z^{LL}$ and $z^{HH} = 0$ in the IC conditions (A.1) and (A.3) with equalities (by $A, C > 0$), obtain:

$$W = \frac{(1 - p_1 + p_0)c}{p_0(p_1 - p_0)} \quad \text{and} \quad z^{LH} = \frac{c}{p_0(p_1 - p_0)}.$$

This contradicts $W - z^{LH} > 0$, so there is no solution in this case.

(c) $p_1 + p_0 < 1$. This implies

$$F = A(1 - p_1) + C(1 - p_1 - p_0) + J \; > \; 0 \; \Rightarrow \; z^{LH} \; = \; 0$$

Using $z^{LH} = z^{HH} = 0$ in the IC conditions (A.1) and (A.3) with equalities, obtain:

$$W = z^{LL} = \frac{(1 + p_1 - p_0)c}{(1 - p_0)(p_1 - p_0)} \quad \text{and} \quad z^{HL} = \frac{c}{1 - p_0}.$$

Using these values in (A.2) leads to a contradiction, so there is no solution in this case.

v. $B > 0\,,\; C > 0\,,\; A = 0$. This implies $D = Bp_1 + C(p_1 + p_0)$.

(a) $p_1 + p_0 = 1$. This implies

$$\begin{cases} E = B(1 - p_1) + I \; > \; 0 \; \Rightarrow \; z^{HL} \; = \; 0 \\ J = Bp_1 + F \; > \; 0 \; \Rightarrow \; W = z^{LH}. \end{cases}$$

Using these values in (A.1) leads to a contradiction, so there is no solution in this case.

(b) $p_1 + p_0 > 1$. This implies

$$J = Bp_1 + C(p_1 + p_0 - 1) + F \; > \; 0 \; \Rightarrow \; W = z^{LH}.$$

Using $W = z^{LH} = z^{LL}, z^{HH} = 0$ in the IC conditions (A.2) and (A.3) with equalities (by $B, C > 0$), obtain:

$$W = \frac{(1 - p_1 + p_0)c}{p_0(p_1 - p_0)} \quad \text{and} \quad z^{HL} = \frac{c}{p_0(p_1 - p_0)},$$

which contradict $W - z^{HL} \geq 0$. So there is no solution in this case.

(c) $p_1 + p_0 < 1$. This implies

$$E = B(1 - p_1) + C(1 - p_1 - p_0) + I \; > \; 0 \;\; \Rightarrow \; z^{HL} \; = \; 0.$$

Using $z^{HL} = z^{HH} = 0$ in the IC conditions (A.2) and (A.3) with equalities, obtain:

$$W = z^{LL} = \frac{(1 + p_1 - p_0)c}{(1 - p_0)(p_1 - p_0)} \quad \text{and} \quad z^{LH} = \frac{c}{1 - p_0},$$

which, when used in (A.1), leads to a contradiction. Thus, there is no solution in this case.

vi. $A > 0\,,\; B > 0\,,\; C = 0$. This implies $D = Ap_1 + Bp_1$ and $D + E = B + I$.

(a) $E > 0$ and $I > 0$. This leads to a contradiction as shown earlier.

(b) $E = 0$ and $I > 0$. This implies $W = z^{HL}$. Insert $W = z^{HL} = z^{LL}, z^{HH} = 0$ into (A.1) and (A.2) with equality (by $A, B > 0$) and we obtain:

$$W = z^{HL} = z^{LL} = \frac{c}{p_1(p_1 - p_0)} \quad \text{and} \quad z^{LH} = \frac{c}{p_1(p_1 - p_0)},$$

which values contradicts (A.3). Hence, there is no solution in this case.

(c) $E > 0$ and $I = 0$. This implies $z^{HL} = 0$. Insert $z^{HL} = z^{HH} = 0$ into (A.1) and (A.2) which are binding due to $A, B > 0$, and we obtain:

$$(1-p_1)(z^{LL}-z^{LH}) = \frac{c}{p_1 - p_0} \quad \text{and} \quad p_1 z^{LH}+(1-p_1)z^{LL} = \frac{c}{p_1 - p_0}.$$

The two equations lead to

$$W = z^{LL} = \frac{c}{(1-p_1)(p_1-p_0)} \quad \text{and} \quad z^{LH} = z^{HL} = z^{HH} = 0.$$

Since $W - z^{LH} > 0$, we have $J = 0$. Therefore, $D + F = A$ and $D + E = B$. Combined with $D + E + F = 1$ and $D = Ap_1 + Bp_1$, we obtain $D = \frac{p_1}{1-p_1}$. Then

$$E + F \;=\; 1 - D \;=\; 1 - \frac{p_1}{1-p_1} \;>\; 0,$$

which implies $p_1 \;<\; \frac{1}{2}$.

Therefore, we have the **solution** for the case of $p_1 < \frac{1}{2}$:

$$\begin{cases} W = \dfrac{c}{(1-p_1)(p_1-p_0)} \\ z^{HH} = z^{HL} = z^{LH} = 0\,,\; z^{LL} = \dfrac{c}{(1-p_1)(p_1-p_0)}. \end{cases}$$

(d) $E = 0$ and $I = 0$. This implies $D = B > 0$. Therefore, we have $B = Ap_1 + Bp_1$ and $B + F = A + J$, which imply $J = \frac{2p_1-1}{p_1}B + F$. We will discuss the different cases of p_1.

- $p_1 = \frac{1}{2}$, so that $A, B > 0$. By inserting $z^{HH} = 0$ into (A.1) and (A.2), we obtain:

$$\frac{1}{2}z^{HL}+\frac{1}{2}(z^{LL}-z^{LH}) = \frac{c}{p_1-p_0} \quad \text{and} \quad \frac{1}{2}z^{LH}+\frac{1}{2}(z^{LL}-z^{HL}) = \frac{c}{p_1-p_0},$$

which leads to

$$W = z^{LL} = \frac{2c}{p_1 - p_0} \quad \text{and} \quad z^{HL} = z^{LH}.$$

By (A.3), we have z^{HL} and $z^{LH} \leq \frac{c}{p_1 - p_0}$. Therefore, we have the **solution** for $p_1 = \frac{1}{2}$:

$$\begin{cases} W = \dfrac{2c}{p_1 - p_0} \\ z^{HH} = 0\,,\, z^{LH} = z^{HL} \in [0, \dfrac{c}{p_1 - p_0}]\,,\, z^{LL} = \dfrac{2c}{p_1 - p_0}. \end{cases}$$

- $p_1 < \frac{1}{2}$, so that $F = J + \frac{1-2p_1}{p_1} B > 0$ and $z^{LH} = 0$. Using $W = z^{LL}$ and $z^{HH} = z^{LH} = 0$ in the IC conditions (A.1) and (A.2) with equalities, obtain

$$p_1 z^{HL} + (1-p_1)(W-0) = \frac{c}{p_1 - p_0} \quad \text{and} \quad (1-p_1)(W - z^{HL}) = \frac{c}{p_1 - p_0}.$$

Thus,

$$W = z^{LL} = \frac{c}{(1-p_1)(p_1 - p_0)} \quad \text{and} \quad z^{HL} = 0.$$

Check (A.3) and other constraints, and we will have the **solution** for $p_1 < \frac{1}{2}$:

$$\begin{cases} W = \dfrac{c}{(1-p_1)(p_1 - p_0)} \\ z^{HH} = z^{HL} = z^{LH} = 0\,,\ z^{LL} = \dfrac{c}{(1-p_1)(p_1 - p_0)}. \end{cases}$$

- $p_1 > \frac{1}{2}$, so that $J > 0$ and $W = z^{LH}$. By inserting $W = z^{LH} = z^{LL}$ and $z^{HH} = 0$ into (A.1) and (A.2) with equality, we have:

$$W = z^{LH} = z^{LL} = z^{HL} = \frac{c}{p_1(p_1 - p_0)}.$$

These values contradict (A.3), so that there is no solution in this case.

vii. $A > 0\,,\; B > 0\,,\; C > 0$. This implies that (A.1)-(A.3) are binding, so that

$$\begin{cases} p_1 z^{HL} + (1-p_1)(z^{LL} - z^{LH}) = \dfrac{c}{p_1 - p_0} \\ p_1 z^{LH} + (1-p_1)(z^{LL} - z^{HL}) = \dfrac{c}{p_1 - p_0} \\ p_1 z^{HL} + (1-p_1)(z^{LL} - z^{LH}) + p_0 z^{LH} + (1-p_0)(z^{LL} - z^{HL}) = \dfrac{2c}{p_1 - p_0}. \end{cases}$$

Therefore, $\quad W = z^{LL} = \dfrac{2c}{p_1 - p_0}$ and $z^{HL} = z^{LH} = \dfrac{c}{p_1 - p_0}$.

Since $W > z^{HL} = z^{LH} > 0$, which implies $W - z^{HL}$, $W - z^{LH}$, z^{HL}, $z^{LH} > 0$, we must have $E = F = I = J = 0$. Then

$$D = K = 1, \qquad A + C = 1$$
$$Ap_1 + Bp_1 + C(p_1 + p_0) = 1, \quad B + C = 1.$$

These imply

$$\begin{cases} C = \frac{2p_1 - 1}{p_1 - p_0} > 0 & \Rightarrow \quad p_1 > \frac{1}{2} \\ A + B = 2(1 - \frac{2p_1 - 1}{p_1 - p_0}) > 0 & \Rightarrow \quad p_1 + p_0 < 1. \end{cases}$$

Therefore, the **solution** for the case of $p_1 > \frac{1}{2}$ and $p_1 + p_0 < 1$ is:

$$\begin{cases} W = \dfrac{2c}{p_1 - p_0} \\ z^{HH} = 0\,,\; z^{HL} = z^{LH} = \dfrac{c}{p_1 - p_0}\,,\; z^{LL} = \dfrac{2c}{p_1 - p_0}. \end{cases}$$

Summarizing the discussions, we have the **solution** for the principal's problem as follows:

(I) If $p_1 > \frac{1}{2}$ and $p_1 + p_0 > 1$,

$$W = \frac{2c}{(p_1+p_0)(p_1-p_0)}\,,\; z^{HH} = 0\,,\; z^{LH} = z^{HL} = z^{LL} = \frac{2c}{(p_1+p_0)(p_1-p_0)}.$$

(II) a. If $p_1 > \frac{1}{2}$ and $p_1 + p_0 = 1$,

$W = \frac{2c}{p_1-p_0}\,,\; z^{HH} = 0\,,\; z^{LL} = \frac{2c}{p_1-p_0}\,,$

$z^{LH}, z^{HL} \in [\frac{c}{p_1-p_0}, \frac{2c}{p_1-p_0}]$ and $p_1 z^{HL} \geq p_0 z^{LH} + c\,,\; p_0 z^{HL} \geq p_1 z^{LH} - c\,.$

b. If $p_1 > \frac{1}{2}$ and $p_1 + p_0 < 1$,

$W = \frac{2c}{p_1-p_0}\,,\; z^{HH} = 0\,,\; z^{LH} = z^{HL} = \frac{c}{p_1-p_0}\,,\; z^{LL} = \frac{2c}{p_1-p_0}.$

c. If $p_1 = \frac{1}{2}$,

$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\; z^{HH} = 0\,,\; z^{LH} = z^{HL} \in [0, \frac{c}{p_1-p_0}]\,,\; z^{LL} = \frac{2c}{p_1-p_0}.$

(III) If $p_1 < \frac{1}{2}$,

$$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\; z^{HH} = z^{LH} = z^{HL} = 0\,,\; z^{LL} = \frac{c}{(1-p_1)(p_1-p_0)}.$$

Q.E.D.

Proof of Proposition 3. We first write the agent's IC conditions.

Period 2. For the agent to exert effort rather than shirk, we require that the continuation value in the second period from effort is no less than from shirking for both first-period states:

$$\begin{cases} W - z_1^H - p_1 z_2^{HH} - (1-p_1) z_2^{HL} - c \geq W - z_1^H - p_0 z_2^{HH} - (1-p_0) z_2^{HL} \\ W - z_1^L - p_1 z_2^{LH} - (1-p_1) z_2^{LL} - c \geq W - z_1^L - p_0 z_2^{LH} - (1-p_0) z_2^{LL}. \end{cases}$$

Period 1. Given that effort will be induced in period two, we consider the following first-period incentive for the agent:

$$\begin{aligned} & V(1,1) \geq V(0,1) \\ \Rightarrow & W - \left\{ p_1 \left[z_1^H + p_1 z_2^{HH} + (1-p_1) z_2^{HL} \right] + (1-p_1) \left[z_1^L + p_1 z_2^{LH} + (1-p_1) z_2^{LL} \right] \right\} - 2c \\ & \geq W - \left\{ p_0 \left[z_1^H + p_1 z_2^{HH} + (1-p_1) z_2^{HL} \right] + (1-p_0) \left[z_1^L + p_1 z_2^{LH} + (1-p_1) z_2^{LL} \right] \right\} - c. \end{aligned}$$

The above IC conditions can be rewritten as:

$$(z_1^L - z_1^H) + p_1(z_2^{LH} - z_2^{HH}) + (1-p_1)(z_2^{LL} - z_2^{HL}) \geq \frac{c}{p_1 - p_0} \tag{A.5}$$

$$z_2^{HL} - z_2^{HH} \geq \frac{c}{p_1 - p_0} \tag{A.6}$$

$$z_2^{LL} - z_2^{LH} \geq \frac{c}{p_1 - p_0}. \tag{A.7}$$

The principal's problem can be written as:

$$\min_{\{W, z_1^H, z_1^L, z_2^{HH}, z_2^{HL}, z_2^{LH}, z_2^{LL}\}} \quad W$$

$$\begin{aligned} s.t. \qquad & \text{(A.5), (A.6), (A.7)} \\ & W - z_1^H - z_2^{HH}, W - z_1^H - z_2^{HL}, W - z_1^L - z_2^{LH}, W - z_1^L - z_2^{LL} \geq 0 \\ & z_1^H, z_1^L, z_2^{HH}, z_2^{LH} \geq 0. \end{aligned}$$

Write the Lagrangian:

$$\begin{aligned} L = & -W + A\left[(z_1^L - z_1^H) + p_1(z_2^{LH} - z_2^{HH}) + (1-p_1)(z_2^{LL} - z_2^{HL}) - \frac{c}{p_1 - p_0}\right] \\ & + B\left[z_2^{HL} - z_2^{HH} - \frac{c}{p_1 - p_0}\right] + C\left[z_2^{LL} - z_2^{LH} - \frac{c}{p_1 - p_0}\right] \\ & + D(W - z_1^H - z_2^{HH}) + E(W - z_1^H - z_2^{HL}) + F(W - z_1^L - z_2^{LH}) + G(W - z_1^L - z_2^{LL}) \\ & + Hz_1^H + Iz_1^L + Jz_2^{HH} + Kz_2^{LH}. \end{aligned}$$

Write the first-order conditions:

$$\frac{\partial L}{\partial W} = -1 + D + E + F + G = 0 \tag{A.8a}$$

$$\frac{\partial L}{\partial z_1^H} = -A - D - E + H = 0 \tag{A.8b}$$

$$\frac{\partial L}{\partial z_1^L} = A - F - G + I = 0 \tag{A.8c}$$

$$\frac{\partial L}{\partial z_2^{HH}} = -Ap_1 - B - D + J = 0 \tag{A.8d}$$

$$\frac{\partial L}{\partial z_2^{HL}} = -A(1-p_1) + B - E = 0 \tag{A.8e}$$

$$\frac{\partial L}{\partial z_2^{LH}} = Ap_1 - C - F + K = 0 \tag{A.8f}$$

$$\frac{\partial L}{\partial z_2^{LL}} = A(1-p_1) + C - G = 0. \tag{A.8g}$$

From (B.4b) and (B.4c), $D + E + F + G = H + I$. By (B.4a), we have

$$H + I = 1. \tag{A.9}$$

From (B.4d) and (B.4e), $-A - D - E + J = 0$. By (B.4b), we have $H = J$. Similarly, by (A.8f), (A.8g) and (B.4c), we have $I = K$. Since $H, I \geq 0$, we can discuss different cases based on (A.9).

i. $H = 1\,, I = 0$. This implies $H = J = 1 > 0$ and $I = K = 0$. Then

$$z_1^H = z_2^{HH} = 0.$$

Since $W > 0$, we have $W - z_1^H - z_2^{HH} > 0$ and $D = 0$. Therefore,

$$\begin{cases} A + E = 1, & Ap_1 + B = 1 \\ A = F + G, & Ap_1 = C + F \\ E + F + G = 1. & \end{cases}$$

Now consider $A + E = 1$. Since $A, E \geq 0$, we have the following possibilities:

(a) $A = 0, E = 1$. This leads to $B = 1$ and $C = F = G = 0$. By $B = E = 1 > 0$ and $z_1^H = z_2^{HH} = 0$, the binding constraint will lead to:

$$W = z_2^{HL} = \frac{c}{p_1 - p_0}.$$

Since

$$\begin{cases} z_2^{LL} \geq z_2^{LH} + \frac{c}{p_1 - p_0} \geq \frac{c}{p_1 - p_0} & \text{by (A.7)} \\ z_2^{LL} \leq W - z_1^L \leq W = \frac{c}{p_1 - p_0} & \text{by } W - z_1^L - z_2^{LL} \geq 0\,, \end{cases}$$

therefore

$$z_2^{LL} = \frac{c}{p_1 - p_0} \text{ and } z_1^L = z_2^{LH} = 0.$$

This turns out to be contradicting with (A.5), so there is no solution in this case.

(b) $A = 1, E = 0$. This implies $B = 1 - p_1 > 0$ ((A.6) is binding). By $z_2^{HH} = 0$,

$$z_2^{HL} = \frac{c}{p_1 - p_0}.$$

Since $F+G = 1$ and $C+F = p_1$, which imply $G = 1-p_1+C > 0$, we have

$$W = z_1^L + z_2^{LL}.$$

From (A.7) , $z_2^{LL} \geq z_2^{LH} + \frac{c}{p_1-p_0} > z_2^{LH}$. Thus

$$W = z_1^L + z_2^{LL} > z_1^L + z_2^{LH},$$

which implies $F = 0$. Then $C = p_1 > 0$, and (A.7) is binding. Then we have

$$z_2^{LL} = z_2^{LH} + \frac{c}{p_1 - p_0} \text{ and } W = z_1^L + z_2^{LL} = z_1^L + z_2^{LH} + \frac{c}{p_1 - p_0}. \tag{A.10}$$

Since $A = 1 > 0$, (A.5) is binding. Inserting $z_1^H = z_2^{HH} = 0$ and $z_2^{HL} = \frac{c}{p_1-p_0}$,

$$\begin{aligned}
\text{LHS of (A.5) } =& p_1(z_1^L + z_2^{LH}) + (1-p_1)(z_1^L + z_2^{LL}) - (1-p_1)\frac{c}{p_1 - p_0} \\
=& z_1^L + z_2^{LL} - p_1\frac{c}{p_1 - p_0} - (1-p_1)\frac{c}{p_1 - p_0} \qquad \text{(by (A.10))} \\
=& W - \frac{c}{p_1 - p_0} \\
=& \text{RHS of (A.5)} = \frac{c}{p_1 - p_0}.
\end{aligned}$$

Therefore,

$$W = \frac{2c}{p_1 - p_0}.$$

Now we have the **solution** for the problem in this case:

$$\begin{aligned}
&W = \frac{2c}{p_1 - p_0} \\
&z_1^H = 0\,,\ z_2^{HH} = 0\,,\ z_2^{HL} = \frac{c}{p_1 - p_0} \\
&z_1^L \in [0, \frac{c}{p_1 - p_0}]\,,\ z_2^{LH} = \frac{c}{p_1 - p_0} - z_1^L\,,\ z_2^{LL} = \frac{2c}{p_1 - p_0} - z_1^L.
\end{aligned}$$

(c) $A > 0, E > 0$. This implies $0 < A < 1$, and so $B > 0$. With both $B, E > 0$, the same contradiction as in (a) applies. Therefore, there is no solution in this case either.

ii. $H = 0\,,\, I = 1$. This implies $H = J = 0$ and $I = K = 1 > 0$. Then

$$z_1^L = z_2^{LH} = 0\,,$$

which also implies $F = 0$. Therefore

$$\begin{cases}
A + D + E = 0, & A + 1 = G \\
Ap_1 + B + D = 0, & Ap_1 + 1 = C \\
D + E + G = 1.
\end{cases}$$

Then we have $A = B = D = E = 0$ and $C = G = 1 > 0$, which implies

$$\begin{cases}
z_2^{LL} = z_2^{LH} + \dfrac{c}{p_1 - p_0} = \dfrac{c}{p_1 - p_0} \\
W = z_1^L + z_2^{LL} = z_2^{LL} = \dfrac{c}{p_1 - p_0}.
\end{cases} \qquad \text{(by } z_1^L = z_2^{LH} = 0\text{)}$$

Since

$$\begin{cases} z_2^{HL} \geq z_2^{HH} + \frac{c}{p_1-p_0} \geq \frac{c}{p_1-p_0} & \text{by (A.6)} \\ z_2^{HL} \leq W - z_1^H \leq W = \frac{c}{p_1-p_0} & \text{by } W - z_1^H - z_2^{HL} \geq 0\,, \end{cases}$$

we have

$$z_2^{HL} = \frac{c}{p_1 - p_0}.$$

However, these values lead to contradiction with (A.5), so that there is no solution in this case.

iii. $H > 0\,,\, I > 0$, this implies $H, I, J, K \in (0,1)$ and it must be

$$z_1^H, z_1^L, z_2^{HH}, z_2^{LH} = 0.$$

These imply that $D = F = 0$ (since $W > 0$), and then $C = Ap_1 + K > 0$, i.e. (A.7) is binding:

$$z_2^{LL} = z_2^{LH} + \frac{c}{p_1 - p_0} = \frac{c}{p_1 - p_0}.$$

These values lead to contradiction with (A.5) again, and there is no solution in this case either.

Summarizing the above discussion, we have the solution for the principal's problem with interim money burning as follows:

$$W = \frac{2c}{p_1 - p_0}$$
$$z_1^H = 0\,,\; z_2^{HH} = 0\,,\; z_2^{HL} = \frac{c}{p_1 - p_0}$$
$$z_1^L \in [0, \frac{c}{p_1 - p_0}]\,,\; z_2^{LH} = \frac{c}{p_1 - p_0} - z_1^L\,,\; z_2^{LL} = \frac{2c}{p_1 - p_0} - z_1^L.$$

Q.E.D.

Proof of Proposition 4.

(i) When $p_1 > \frac{1}{2}$ and $p_1 + p_0 \leq 1$, the optimal contracts are the same for the two scenarios.

(ii) When $p_1 > \frac{1}{2}$ and $p_1 + p_0 > 1$, $\frac{2c}{(p_1+p_0)(p_1-p_0)} < \frac{2c}{p_1-p_0}$.

(iii) When $p_1 \leq \frac{1}{2}$, $\frac{c}{(1-p_1)(p_1-p_0)} \leq \frac{2c}{p_1-p_0}$.

For the cases (ii) and (iii), LHS is the total budget for the principal without interim money burning (see Proposition 1), and RHS is the budget with interim money burning (see Proposition 2). It can be seen that the principal saves cost without interim money burning. **Q.E.D.**

Proof of Proposition **??**. Without interim money burning, we can write the incentive compatibility conditions in terms of agent's total cost: $TC(1,1) \leq \min\{TC(1,0)\,, TC(0,1)\,, TC(0,0)\}$. Explicitly,

$$\begin{aligned} &p_1^2 z^{HH} + p_1(1-p_1)z^{HL} + (1-p_1)p_1 z^{LH} + (1-p_1)^2 z^{LL} + 2c \\ &\qquad \leq\; p_1p_0 z^{HH} + p_1(1-p_0)z^{HL} + (1-p_1)p_0 z^{LH} + (1-p_1)(1-p_0)z^{LL} + c\,; \end{aligned} \tag{A.11}$$

$$\begin{aligned} &p_1^2 z^{HH} + p_1(1-p_1)z^{HL} + (1-p_1)p_1 z^{LH} + (1-p_1)^2 z^{LL} + 2c \\ &\qquad \leq\; p_0p_1 z^{HH} + p_0(1-p_1)z^{HL} + (1-p_0)p_1 z^{LH} + (1-p_0)(1-p_1)z^{LL} + c\,; \end{aligned} \tag{A.12}$$

$$\begin{aligned} &p_1^2 z^{HH} + p_1(1-p_1)z^{HL} + (1-p_1)p_1 z^{LH} + (1-p_1)^2 z^{LL} + 2c \\ &\qquad \leq\; p_0p_0 z^{HH} + p_0(1-p_0)z^{HL} + (1-p_0)p_0 z^{LH} + (1-p_0)(1-p_0)z^{LL}\,. \end{aligned} \tag{A.13}$$

With interim money burning, the continuation incentive compatibility conditions can be written as:

$$p_1 z_2^{HH} + (1-p_1)z_2^{HL} + c \;\leq\; p_0 z_2^{HH} + (1-p_0)z_2^{HL} \tag{A.14}$$

$$p_1 z_2^{LH} + (1-p_1)z_2^{LL} + c \;\leq\; p_0 z_2^{LH} + (1-p_0)z_2^{LL}. \tag{A.15}$$

Given effort in period two, we have the first-period constraint:

$$TC(1,1) \le TC(0,1)$$

i.e., $p_1 z_1^H + (1-p_1)z_1^L + p_1\left[p_1 z_2^{HH} + (1-p_1)z_2^{HL}\right] + (1-p_1)\left[p_1 z_2^{LH} + (1-p_1)z_2^{LL}\right] + 2c$

$$\le p_0 z_1^H + (1-p_0)z_1^L + p_0\left[p_1 z_2^{HH} + (1-p_1)z_2^{HL}\right] + (1-p_0)\left[p_1 z_2^{LH} + (1-p_1)z_2^{LL}\right] + c. \tag{A.16}$$

Let

$$\begin{aligned} z^{HH} &= z_1^H + z_2^{HH}\,,\ z^{HL} = z_1^H + z_2^{HL}\,, \\ z^{LH} &= z_1^L + z_2^{LH}\,,\ z^{LL} = z_1^L + z_2^{LL}\,, \end{aligned} \tag{A.17}$$

we will show that any $\{z_1^H,\, z_1^L,\, z_2^{HH},\, z_2^{HL},\, z_2^{LH},\, z_2^{LL}\}$ that satisfy (A.14)–(A.16) imply (A.11)–(A.13).

i. By (A.16),

LHS of (A.16)

$= p_1\left[p_1(z_1^H + z_2^{HH}) + (1-p_1)(z_1^H + z_2^{HL})\right] + (1-p_1)\left[p_1(z_1^L + z_2^{LH}) + (1-p_1)(z_1^L + z_2^{LL})\right] + 2c$

$= p_1^2 z^{HH} + p_1(1-p_1)z^{HL} + (1-p_1)p_1 z^{LH} + (1-p_1)^2 z^{LL} + 2c =$ **LHS of** (A.12)

$\le$ RHS of (A.16)

$= p_0\left[p_1(z_1^H + z_2^{HH}) + (1-p_1)(z_1^H + z_2^{HL})\right] + (1-p_0)\left[p_1(z_1^L + z_2^{LH}) + (1-p_1)(z_1^L + z_2^{LL})\right] + c$

$= p_0 p_1 z^{HH} + p_0(1-p_1)z^{HL} + (1-p_0)p_1 z^{LH} + (1-p_0)(1-p_1)z^{LL} + c =$ **RHS of** (A.12).

By the above series of inequalities we have shown that (A.16) implies (A.12).

ii. By (A.14) and (A.15),

$$\begin{aligned} &p_1 z_1^H + (1-p_1)z_1^L + c + p_1\left[p_1 z_2^{HH} + (1-p_1)z_2^{HL} + c\right] + (1-p_1)\left[p_1 z_2^{LH} + (1-p_1)z_2^{LL} + c\right] \\ &\quad \le p_1 z_1^H + (1-p_1)z_1^L + c + p_1\left[p_0 z_2^{HH} + (1-p_0)z_2^{HL}\right] + (1-p_1)\left[p_0 z_2^{LH} - (1-p_0)z_2^{LL}\right], \end{aligned}$$

which can be rewritten as:

$$\begin{aligned}
&p_1\left[p_1(z_1^H+z_2^{HH})+(1-p_1)(z_1^H+z_2^{HL})\right]+(1-p_1)\left[p_1(z_1^L+z_2^{LH})+(1-p_1)(z_1^L+z_2^{LL})\right]+2c\\
&\quad\le p_1\left[p_0(z_1^H+z_2^{HH})+(1-p_0)(z_1^H+z_2^{HL})\right]+(1-p_1)\left[p_0(z_1^L+z_2^{LH})+(1-p_0)(z_1^L+z_2^{LL})\right]+c\\
&\Leftrightarrow\\
&p_1^2z^{HH}+p_1(1-p_1)z^{HL}+(1-p_1)p_1z^{LH}+(1-p_1)^2z^{LL}+2c\\
&\quad\le p_1p_0z^{HH}+p_1(1-p_0)z^{HL}+(1-p_1)p_0z^{LH}+(1-p_1)(1-p_0)z^{LL}+c.
\end{aligned}$$

By this we have shown that (A.14) and (A.15) imply (A.11).

iii. Similarly, by (A.14) and (A.15),

$$\begin{aligned}
&p_0z_1^H+(1-p_0)z_1^L+p_0\left[p_1z_2^{HH}+(1-p_1)z_2^{HL}+c\right]+(1-p_0)\left[p_1z_2^{LH}+(1-p_1)z_2^{LL}+c\right]\\
&\quad\le p_0z_1^H+(1-p_0)z_1^L+p_0\left[p_0z_2^{HH}+(1-p_0)z_2^{HL}\right]+(1-p_0)\left[p_0z_2^{LH}-(1-p_0)z_2^{LL}\right]
\end{aligned}$$

which can be rewritten as:

$$\begin{aligned}
&p_0\left[p_1(z_1^H+z_2^{HH})+(1-p_1)(z_1^H+z_2^{HL})\right]+(1-p_0)\left[p_1(z_1^L+z_2^{LH})+(1-p_1)(z_1^L+z_2^{LL})\right]+c\\
&\quad\le p_0\left[p_0(z_1^H+z_2^{HH})+(1-p_0)(z_1^H+z_2^{HL})\right]+(1-p_0)\left[p_0(z_1^L+z_2^{LH})+(1-p_0)(z_1^L+z_2^{LL})\right]
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow\ p_0p_1z^{HH}+p_0(1-p_1)z^{HL}+(1-p_0)p_1z^{LH}+(1-p_0)(1-p_1)z^{LL}+c\\
&\quad\le p_0^2z^{HH}+p_0(1-p_0)z^{HL}+(1-p_0)p_0z^{LH}+(1-p_0)^2z^{LL}.
\end{aligned}$$

By (A.12),

$$\begin{aligned}
&p_1^2z^{HH}+p_1(1-p_1)z^{HL}+(1-p_1)p_1z^{LH}+(1-p_1)^2z^{LL}+2c\\
&\quad\le p_0^2z^{HH}+p_0(1-p_0)z^{HL}+(1-p_0)p_0z^{LH}+(1-p_0)^2z^{LL}.
\end{aligned}$$

Therefore, (A.14), (A.15) and (A.12) imply (A.13).

In conclusion, by manipulating the money burning terms as in (A.17), any contract with interim money burning can be replicated by a contract without interim money burning (in terms of incentives). **Q.E.D.**

APPENDIX B

Proofs for Chapter 2

Proof of Proposition 5. The agent's payoffs in the repeated efforts game are as follows:

$$V(1,1) = W - p_1p_1z^{HH} - p_1(1-p_1)z^{HL} - (1-p_1)p_1z^{LH} - (1-p_1)(1-p_1)z^{LL} - 2c$$
$$V(1,0) = W - p_1p_0z^{HH} - p_1(1-p_0)z^{HL} - (1-p_0)p_1z^{LH} - (1-p_1)(1-p_0)z^{LL} - c$$
$$V(0,1) = W - p_0p_1z^{HH} - p_0(1-p_1)z^{HL} - (1-p_1)p_0z^{LH} - (1-p_0)(1-p_1)z^{LL} - c$$
$$V(0,0) = W - p_0p_0z^{HH} - p_0(1-p_0)z^{HL} - (1-p_0)p_0z^{LH} - (1-p_0)(1-p_0)z^{LL}.$$

For full efforts implementation, the agent's incentive compatibility (IC) conditions are:

$$p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) \geq \frac{c}{p_1 - p_0} \tag{B.1}$$
$$p_1(z^{LH} - z^{HH}) + (1-p_1)(z^{LL} - z^{HL}) \geq \frac{c}{p_1 - p_0} \tag{B.2}$$
$$p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) + p_0(z^{LH} - z^{HH}) + (1-p_0)(z^{LL} - z^{HL}) \geq \frac{2c}{p_1 - p_0}. \tag{B.3}$$

Thus, the principal's problem can be written as follows:

$$\min_{\{W, z^{HH}, z^{HL}, z^{LH}, z^{LL}\}} \quad W$$

$$\begin{aligned} s.t. \quad & \text{(A.1)}, \text{(A.2)}, \text{(A.3)} \\ & z^{HH}, z^{HL}, z^{LH}, z^{LL} \geq 0 \\ & W - z^{HH}, W - z^{HL}, W - z^{LH}, W - z^{LL} \geq 0. \end{aligned}$$

Write the Lagrangian:

$$\begin{aligned} L = & -W + A\left[p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) - \frac{c}{p_1 - p_0}\right] \\ & + B\left[p_1(z^{LH} - z^{HH}) + (1-p_1)(z^{LL} - z^{HL}) - \frac{c}{p_1 - p_0}\right] \\ & + C\left[p_1(z^{HL} - z^{HH}) + (1-p_1)(z^{LL} - z^{LH}) + p_0(z^{LH} - z^{HH}) + (1-p_0)(z^{LL} - z^{HL}) - \frac{2c}{p_1 - p_0}\right] \\ & + Dz^{HH} + Ez^{HL} + Fz^{LH} + Gz^{LL} \\ & + H(W - z^{HH}) + I(W - z^{HL}) + J(W - z^{LH}) + K(W - z^{LL}). \end{aligned}$$

FOCs are:

$$\frac{\partial L}{\partial W} = -1 + H + I + J + K = 0 \tag{B.4a}$$

$$\frac{\partial L}{\partial z^{HH}} = -Ap_1 - Bp_1 - Cp_1 - Cp_0 + D - H = 0 \tag{B.4b}$$

$$\frac{\partial L}{\partial z^{HL}} = Ap_1 - B(1-p_1) + Cp_1 - C(1-p_0) + E - I = 0 \tag{B.4c}$$

$$\frac{\partial L}{\partial z^{LH}} = -A(1-p_1) + Bp_1 - C(1-p_1) + Cp_0 + F - J = 0 \tag{B.4d}$$

$$\frac{\partial L}{\partial z^{LL}} = A(1-p_1) + B(1-p_1) + C(1-p_1) + C(1-p_0) + G - K = 0. \tag{B.4e}$$

Applying the complementary slackness conditions to FOCs (Kuhn-Tucker method), we obtain the following solution :

(I) If $p_1 > \frac{1}{2}$ and $p_1 + p_0 > 1$,

$$W = \frac{2c}{(p_1+p_0)(p_1-p_0)}\,,\ z^{HH} = 0\,,\ z^{LH} = z^{HL} = z^{LL} = \frac{2c}{(p_1+p_0)(p_1-p_0)}.$$

(II) a. If $p_1 > \frac{1}{2}$ and $p_1 + p_0 = 1$,

$W = \frac{2c}{p_1 - p_0}\,,\ z^{HH} = 0\,,\ z^{LL} = \frac{2c}{p_1 - p_0}\,,$

$z^{LH}, z^{HL} \in [\frac{c}{p_1-p_0}, \frac{2c}{p_1-p_0}]$ and $p_1 z^{HL} \geq p_0 z^{LH} + c\,,\ p_0 z^{HL} \geq p_1 z^{LH} - c\,.$

b. If $p_1 > \frac{1}{2}$ and $p_1 + p_0 < 1$,

$W = \frac{2c}{p_1 - p_0}\,,\ z^{HH} = 0\,,\ z^{LH} = z^{HL} = \frac{c}{p_1 - p_0}\,,\ z^{LL} = \frac{2c}{p_1 - p_0}.$

c. If $p_1 = \frac{1}{2}$,

$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\ z^{HH} = 0\,,\ z^{LH} = z^{HL} \in [0, \frac{c}{p_1-p_0}]\,,\ z^{LL} = \frac{2c}{p_1-p_0}.$

(III) If $p_1 < \frac{1}{2}$,

$$W = \frac{c}{(1-p_1)(p_1-p_0)}\,,\ z^{HH} = z^{LH} = z^{HL} = 0\,,\ z^{LL} = \frac{c}{(1-p_1)(p_1-p_0)}.$$

Q.E.D.

DERIVATION OF TABLE 2.2: *Optimal mechanism with agent's information.* We first analyze the incentive compatibility conditions using backward induction.

Period 2. We need to impose money burnings such that the agent is to exert effort in the second period no matter what private signal he receives for the first period, i.e., in both information sets $(1, s_G)$ and $(1, s_B)$:

$$\begin{aligned} W - & q^{1G}[p_1 z^{HH} + (1-p_1) z^{HL}] - (1-q^{1G})[p_1 z^{LH} + (1-p_1) z^{LL}] - 2c \\ & \geq W - q^{1G}[p_0 z^{HH} + (1-p_0) z^{HL}] - (1-q^{1G})[p_0 z^{LH} + (1-p_0) z^{LL}] - c \\ W - & q^{1B}[p_1 z^{HH} + (1-p_1) z^{HL}] - (1-q^{1B})[p_1 z^{LH} + (1-p_1) z^{LL}] - 2c \\ & \geq W - q^{1B}[p_0 z^{HH} + (1-p_0) z^{HL}] - (1-q^{1B})[p_0 z^{LH} + (1-p_0) z^{LL}] - c. \end{aligned}$$

Period 1. Consider the following first-period incentive for the agent:

$V(1,1) \geq V(0,1)$ and $V(1,1) \geq V(0,0),$

i.e.,
$$\begin{aligned}
&W - p_1^2 z^{HH} - p_1(1-p_1)z^{HL} - (1-p_1)p_1 z^{LH} - (1-p_1)^2 z^{LL} - 2c \\
&\quad \geq W - p_0 p_1 z^{HH} - p_0(1-p_1)z^{HL} - (1-p_0)p_1 z^{LH} - (1-p_0)(1-p_1)z^{LL} - c \\
&W - p_1^2 z^{HH} - p_1(1-p_1)z^{HL} - (1-p_1)p_1 z^{LH} - (1-p_1)^2 z^{LL} - 2c \\
&\quad \geq W - p_0 p_0 z^{HH} - p_0(1-p_0)z^{HL} - (1-p_0)p_0 z^{LH} - (1-p_0)(1-p_0)z^{LL}.
\end{aligned}$$

Manipulating the conditions above, we have to distinguish two cases.

■ **Case 1:**

$$z^{LL} - z^{LH} \geq z^{HL} - z^{HH}. \qquad \text{(B.5)}$$

Since $q^{1G} > q^{1B} > q^{0G} > q^{0B}$, given (B.5) it can be shown that once the agent chooses second period effort upon $(1, s_G)$, he will always choose effort upon $(1, s_B)$, $(0, s_G)$ and $(0, s_B)$. Therefore, the incentive compatibility conditions can be simplified to:

$$\begin{aligned}
&\textbf{[IC}_1\textbf{]} \quad q^{1G}(z^{HL} - z^{HH}) + (1-q^{1G})(z^{LL} - z^{LH}) \geq \frac{c}{p_1 - p_0}, \\
&\textbf{[IC}_3\textbf{]} \quad p_1(z^{LH} - z^{HH}) + (1-p_1)(z^{LL} - z^{HL}) \geq \frac{c}{p_1 - p_0}.
\end{aligned}$$

Then the principal's problem can be written as:

$$\begin{aligned}
\min_{W, z^{HH}, z^{HL}, z^{LH}, z^{LL}} \quad W \quad s.t. \quad & [\text{IC}_1], [\text{IC}_3] \\
& W - z^{HH} \geq 0,\ W - z^{HL} \geq 0,\ W - z^{LH} \geq 0,\ W - z^{LL} \geq 0 \\
& z^{HH} \geq 0,\ z^{HL} \geq 0,\ z^{LH} \geq 0,\ z^{LL} \geq 0.
\end{aligned}$$

The Lagrangian is:

$$\begin{aligned}
L = -W &+ A\left[q^{1G}(z^{HL} - z^{HH}) + (1-q^{1G})(z^{LL} - z^{LH}) - \frac{c}{p_1 - p_0}\right] \\
&+ B\left[p_1(z^{LH} - z^{HH}) + (1-p_1)(z^{LL} - z^{HL}) - \frac{c}{p_1 - p_0}\right] \\
&+ C(W - z^{HH}) + D(W - z^{HL}) + E(W - z^{LH}) + F(W - z^{LL}) \\
&+ Gz^{HH} + Hz^{HL} + Iz^{LH} + Jz^{LL}.
\end{aligned}$$

Now the first order conditions can be written as follows:

$$\frac{\partial L}{\partial W} = -1 + C + D + E + F = 0 \tag{B.6a}$$

$$\frac{\partial L}{\partial z^{HH}} = -Aq^{1G} - Bp_1 - C + G = 0 \tag{B.6b}$$

$$\frac{\partial L}{\partial z^{HL}} = Aq^{1G} - B(1-p_1) - D + H = 0 \tag{B.6c}$$

$$\frac{\partial L}{\partial z^{LH}} = -A(1-q^{1G}) + Bp_1 - E + I = 0 \tag{B.6d}$$

$$\frac{\partial L}{\partial z^{LL}} = A(1-q^{1G}) + B(1-p_1) - F + J = 0. \tag{B.6e}$$

From (B.6b), $G = Aq^{1G} + Bp_1 + C$. If $G = 0$, then $A = B = C = 0$, which imply $D = H$, $E = I$, $F = J$ and $D + E + F = 1$. By the last equation, all D, E and F cannot be zero. Suppose $D > 0$, which implies also $H > 0$, then by complementary slackness we have $W - z^{HL} = 0$ and $z^{HL} = 0$. Thus $W = z^{HL} = 0$, which will lead to contradiction to the incentive compatibility conditions. Therefore, it must be that $D = 0$. The same argument applies to E and F, so that it cannot be $D + E + F = 1 > 0$, and thus $G = 0$ should not hold. Hence we have $G > 0$, and then $z^{HH} = 0$. By a similar analysis, F can be determined to be positive, which tells that $W - z^{LL} = 0$ and thus $W = z^{LL}$. Now the FOCs can be simplified to:

$$\begin{cases} D + E + F = 1 \\ Aq^{1G} - B(1-p_1) - D + H = 0\,, \quad A(1-q^{1G}) - Bp_1 + E - I = 0 \\ G = Aq^{1G} + Bp_1 > 0\,, \quad F = A(1-q^{1G}) + B(1-p_1) > 0. \end{cases}$$

Since $Aq^{1G} + Bp_1 > 0$, both A and B cannot be zero. Then we have the following discussions.

- If $A > 0$ and $B = 0$, then $D = Aq^{1G} + H > 0$ and $I = A(1-q^{1G}) + E > 0$. These imply $W = z^{HL}$ and $z^{LH} = 0$. Substituting the values together with $W = z^{LL}$ and $z^{HH} = 0$ into [IC$_3$] leads to a contradiction. So there is no solution in this case.

- If $A = 0$ and $B > 0$, then $H = B(1-p_1) + D > 0$ and $E = Bp_1 +$

$I > 0$. These imply $z^{HL} = 0$ and $W = z^{LH}$. Substituting the values together with $W = z^{LL}$ and $z^{HH} = 0$ into [IC$_1$] leads to a contradiction. So there is also no solution in this case.

- If $A > 0$ and $B > 0$, then both [IC$_1$] and [IC$_3$] are binding, and we have the following equations:

$$q^{1G}z^{HL} + (1-q^{1G})(W - z^{LH}) = \frac{c}{p_1 - p_0} \tag{B.7}$$

$$p_1 z^{LH} + (1-p_1)(W - z^{HL}) = \frac{c}{p_1 - p_0} \tag{B.8}$$

Now we need to discuss D and E.

i. $D > 0$ and $E > 0$. These imply $W = z^{HL} = z^{LH}$. Using the values in (B.7) and (B.8) yields $q^{1G} = p_1$, which is a contradiction to the assumption. So there is no solution in this case.

ii. $D > 0$ and $E = 0$. The former implies $W = z^{HL}$, which can be used in (B.7) and (B.8). Then we obtain $z^{LH} = \frac{c}{p_1(p_1-p_0)}$ and $W = \frac{1-q^{1G}+p_1}{p_1}\frac{c}{p_1-p_0}$, which leads to $W - z^{LH} = \frac{p_1-q^{1G}}{p_1}\frac{c}{p_1-p_0} < 0$, a contradiction. So there is no solution in this case.

iii. $D = 0$ and $E > 0$. The latter implies $W = z^{LH}$. Substituting it into (B.7) and (B.8) gives us $z^{HL} = \frac{c}{q^{1G}(p_1-p_0)}$ and $W = \frac{1+q^{1G}-p_1}{q^{1G}}\frac{c}{p_1-p_0}$. However, now $z^{HL} - z^{HH} = \frac{c}{q^{1G}(p_1-p_0)} > z^{LL} - z^{LH} = 0$, which contradicts (B.5), so there is still no solution in this case.

iv. $D = 0$ and $E = 0$. In this case, if $H > 0$, then $z^{HL} = 0$, which leads to $z^{LH} = \frac{p_1-q^{1G}}{1-q^{1G}}\frac{c}{p_1-p_0} < 0$ after substituting into (B.7) and (B.8). Hence it must be $H = 0$. Then we have $F = 1$ and $Aq^{1G} = B(1-p_1)$, so that $F = A(1-q^{1G}) + B(1-p_1) = A(1-q^{1G}) + Aq^{1G} = A = 1$ and $B = \frac{q^{1G}}{1-p_1}$, $I = A(1-q^{1G}) - Bp_1 = \frac{1-q^{1G}-p_1}{1-p_1}$. Since $I \geq 0$, we must have $p_1 + q^{1G} \leq 1$.

- When $p_1 + q^{1G} < 1$, this implies $I > 0$. Hence $z^{LH} = 0$ and (B.7) and (B.8) can be solved:

$$W = \frac{1 - p_1 + q^{1G}}{1 - p_1} \frac{c}{p_1 - p_0} = \frac{1 + \rho(1 - p_1)}{1 - p_1} \frac{c}{p_1 - p_0}$$
$$z^{HL} = \frac{q^{1G} - p_1}{1 - p_1} \frac{c}{p_1 - p_0} = \rho \frac{c}{p_1 - p_0}.$$

- When $p_1 + q^{1G} = 1$, (B.7) and (B.8) can be written as:

$(1-p_1)z^{HL} + p_1(W - z^{LH}) = \frac{c}{p_1 - p_0}$ and $p_1 z^{LH} + (1-p_1)(W - z^{HL}) = \frac{c}{p_1 - p_0}$,

which gives us

$$W = \frac{2c}{p_1 - p_0} = \frac{2(p_1 + q^{1G})c}{p_1 - p_0} = \frac{1 + \rho(1 - p_1)}{1 - p_1} \frac{c}{p_1 - p_0}.$$

Given a range of money burning values, we pick the one with the lowest expected value:

$$z^{LH} = 0 \quad \text{and} \quad z^{HL} = \frac{(1 - 2p_1)c}{(1 - p_1)(p_1 - p_0)} = \rho \frac{c}{p_1 - p_0}.$$

It can be verified that the solutions derived above satisfy (B.5).

■ **Case 2:**

$$z^{LL} - z^{LH} \leq z^{HL} - z^{HH}. \tag{B.9}$$

Now given (B.9), once agent chooses second period effort upon $(1, s_B)$, he will also choose effort upon $(1, s_B)$. Then the incentive compatibility conditions are:

[IC$_2$] $q^{1B}(z^{HL} - z^{HH}) + (1 - q^{1B})(z^{LL} - z^{LH}) \geq \frac{c}{p_1 - p_0}$,

[IC$_3$] $p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) \geq \frac{c}{p_1 - p_0}$,

[IC$_4$] $p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL})$
$\quad + p_0(z^{HL} - z^{HH}) + (1 - p_0)(z^{LL} - z^{LH}) \geq \frac{2c}{p_1 - p_0}$.

Then the principal's problem can be written as:

$$\min_{W,z^{HH},z^{HL},z^{LH},z^{LL}} \quad W \quad s.t. \qquad \text{[IC}_2\text{], [IC}_3\text{], [IC}_4\text{]}$$

$$W - z^{HH} \geq 0,\ W - z^{HL} \geq 0,\ W - z^{LH} \geq 0,\ W - z^{LL} \geq 0$$

$$z^{HH} \geq 0,\ z^{HL} \geq 0,\ z^{LH} \geq 0,\ z^{LL} \geq 0.$$

The Lagrangian is:

$$\begin{aligned}
L = & - W + A\left[q^{1B}(z^{HL} - z^{HH}) + (1 - q^{1B})(z^{LL} - z^{LH}) - \frac{c}{p_1 - p_0}\right] \\
& + B\left[p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) - \frac{c}{p_1 - p_0}\right] \\
& + C\left[p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) + p_0(z^{HL} - z^{HH}) + (1 - p_0)(z^{LL} - z^{LH}) - \frac{2c}{p_1 - p_0}\right] \\
& + Dz^{HH} + Ez^{HL} + Fz^{LH} + Gz^{LL} \\
& + H(W - z^{HH}) + I(W - z^{HL}) + J(W - z^{LH}) + K(W - z^{LL}).
\end{aligned}$$

Now the first order conditions can be written as follows:

$$\frac{\partial L}{\partial W} = -1 + H + I + J + K = 0 \qquad \text{(B.10a)}$$

$$\frac{\partial L}{\partial z^{HH}} = -Aq^{1B} - Bp_1 - Cp_1 - Cp_0 + D - H = 0 \qquad \text{(B.10b)}$$

$$\frac{\partial L}{\partial z^{HL}} = Aq^{1B} - B(1 - p_1) - C(1 - p_1) + Cp_0 + E - I = 0 \qquad \text{(B.10c)}$$

$$\frac{\partial L}{\partial z^{LH}} = -A(1 - q^{1B}) + Bp_1 + Cp_1 - C(1 - p_0) + F - J = 0 \qquad \text{(B.10d)}$$

$$\frac{\partial L}{\partial z^{LL}} = A(1 - q^{1B}) + B(1 - p_1) + C(1 - p_1) + C(1 - p_0) + G - K = 0. \qquad \text{(B.10e)}$$

Following the similar arguments as in the first case, it can be determined that $D > 0$ and $K > 0$, which imply $z^{HH} = 0$ and $W = z^{LL}$, as well as $H = 0$ and $G = 0$. Then we have $D = Aq^{1B} + Bp_1 + C(p_1 + p_0) > 0$, so that all A, B and C cannot be zero, and the following cases are possible.

i. $A > 0, B = C = 0$. This implies $I = Aq^{1B}+E > 0$, and thus $W = z^{HL}$. Also we have $F = A(1-q^{1B}) + J > 0$, and thus $z^{LH} = 0$. Using these values together with $z^{HH} = 0$ and $W = z^{LL}$ in [IC$_3$], it turns out to be a contradiction. So there is no solution in this case.

ii. $B > 0$, $A = C = 0$. Following a similar argument as in case i above, it leads to a contradiction with [IC$_2$]. So there is also no solution in this case.

iii. $C > 0$, $A = B = 0$. This implies

$$p_1(z^{LH}-z^{HH})+(1-p_1)(z^{LL}-z^{HL})+p_0(z^{HL}-z^{HH})+(1-p_0)(z^{LL}-z^{LH}) = \frac{2c}{p_1-p_0}, \tag{B.11}$$

and $D = C(p_1 + p_0)$. Then we need to discuss different values of $p_1 + p_0$.

(a) If $p_1 + p_0 = 1$, then we can derive from (B.11) that $W = z^{LL} = \frac{2c}{p_1-p_0}$. By (B.9), we have $W - z^{LH} \leq z^{HL} - 0$, i.e., $z^{HL} + z^{LH} \geq \frac{2c}{p_1-p_0}$. Substituting the values of z^{LL} and $z^{HH} = 0$ into [IC$_2$] and [IC$_3$], the following should hold:

$$\begin{cases} q^{1B}z^{HL} - (1-q^{1B})z^{LH} \geq \left[1 - 2(1-q^{1B})\right] \dfrac{c}{p_1-p_0} \\ p_1 z^{LH} - (1-p_1)z^{HL} \geq [1 - 2(1-p_1)] \dfrac{c}{p_1-p_0}. \end{cases}$$

Since $p_1+p_0 = 1$ and $1-q^{1B} = 1-p_1+\rho(1-p_1) = p_0+\rho p_0 = q^{0G}$, the above can be re-written as:

$$\begin{cases} q^{1B}z^{HL} - q^{0G}z^{LH} \geq (q^{1B} - q^{0G})\dfrac{c}{p_1-p_0} \\ p_1 z^{LH} - p_0 z^{HL} \geq (p_1 - p_0)\dfrac{c}{p_1-p_0}. \end{cases}$$

i.e.,
$$\begin{cases} q^{1B}\left(z^{HL} - \dfrac{c}{p_1-p_0}\right) \geq q^{0G}\left(z^{LH} - \dfrac{c}{p_1-p_0}\right) \\ p_1\left(z^{LH} - \dfrac{c}{p_1-p_0}\right) \geq p_0\left(z^{HL} - \dfrac{c}{p_1-p_0}\right). \end{cases}$$

Suppose $z^{HL} - \frac{c}{p_1-p_0} < 0$, then $z^{LH} - \frac{c}{p_1-p_0} < 0$. The above two inequalities lead to:

$$\frac{z^{HL} - \frac{c}{p_1-p_0}}{z^{LH} - \frac{c}{p_1-p_0}} \leq \frac{q^{0G}}{q^{1B}} < 1 \quad \text{and} \quad \frac{z^{HL} - \frac{c}{p_1-p_0}}{z^{LH} - \frac{c}{p_1-p_0}} \geq \frac{p_1}{p_0} > 1,$$

which are contradicting with each other. So it must be $z^{HL} - \frac{c}{p_1-p_0} \geq 0$. By a similar argument, we also have $z^{LH} - \frac{c}{p_1-p_0} \geq 0$. Then picking the smallest values of z^{HL} and z^{LH}, we have

$$z^{HL} = z^{LH} = \frac{c}{p_1 - p_0}.$$

In this case, both [IC$_2$] and [IC$_3$] bind with equalities.

(b) If $p_1 + p_0 > 1$, then we have $I = C(p_1 + p_0 - 1) + E > 0$ and $J = C(p_1 + p_0 - 1) + F > 0$, which imply that $W = z^{HL}$ and $W = z^{LH}$. Since $z^{HH} = 0$ and $W = z^{LL}$, we can now derive from (B.11) the solution as follows:

$$W = z^{HL} = z^{LH} = z^{LL} = \frac{2c}{(p_1 + p_0)(p_1 - p_0)}\ , \quad z^{HH} = 0.$$

It can be verified that [IC$_2$] and [IC$_3$] hold with strict inequality in this case.

(c) If $p_1 + p_0 < 1$, then we have $E = I + C(1 - p_1 - p_0) > 0$ and $F = J + C(1-p_1-p_0) > 0$, which imply that $z^{HL} = 0$ and $z^{LH} = 0$. Since $W = z^{LL}$ and $z^{HH} = 0$, it yields $z^{LL} - z^{LH} > z^{HL} - z^{HH} = 0$, which is a contradiction to (B.9). So there is no solution in this case.

iv. $A > 0$, $B > 0$ and $C = 0$. These imply that

$$\begin{cases} q^{1B}(z^{HL} - z^{HH}) + (1 - q^{1B})(z^{LL} - z^{LH}) = \dfrac{c}{p_1 - p_0} \\ p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) = \dfrac{c}{p_1 - p_0}. \end{cases}$$

Since $q^{1B} > p_0$, by (B.9) we obtain:

$$p_0(z^{HL}-z^{HH})+(1-p_0)(z^{LL}-z^{LH}) < q^{1B}(z^{HL}-z^{HH})+(1-q^{1B})(z^{LL}-z^{LH}) = \frac{c}{p_1 - p_0}.$$

Therefore,

$$p_1(z^{LH}-z^{HH})+(1-p_1)(z^{LL}-z^{HL})+p_0(z^{HL}-z^{HH})+(1-p_0)(z^{LL}-z^{LH}) < \frac{2c}{p_1-p_0},$$

which contradicts [IC$_4$]. So there is no solution in this case.

v. $A > 0$, $B = 0$ and $C > 0$. These imply that

$$\begin{cases} q^{1B}(z^{HL}-z^{HH})+(1-q^{1B})(z^{LL}-z^{LH}) = \dfrac{c}{p_1-p_0} \\ p_1(z^{LH}-z^{HH})+(1-p_1)(z^{LL}-z^{HL})+p_0(z^{HL}-z^{HH})+(1-p_0)(z^{LL}-z^{LH}) = \dfrac{2c}{p_1-p_0}. \end{cases} \tag{B.12}$$

Further, we have $D = Aq^{1B} + C(p_1 + p_0)$. Again we need to discuss different values of $p_1 + p_0$.

(a) If $p_1 + p_0 = 1$, then $I = Aq^{1B} + E > 0$ and $F = A(1-q^{1B}) + J > 0$, which imply $W = z^{HL}$ and $z^{LH} = 0$. Substituting the values into [IC$_3$] yields a contradiction, so that there is no solution in this case.

(b) If $p_1 + p_0 > 1$, then $I = Aq^{1B} + C(p_1 + p_0 - 1) + E > 0$, which implies $W = z^{HL}$. Using the value in (B.12) we obtain:

$$W - z^{LH} = \frac{p_1 + p_0 - 2q^{1B}}{p_1 + p_0 - q^{1B}} \frac{c}{p_1 - p_0}.$$

Since $p_1 + p_0 - 2q^{1B} = p_1 + p_0 - 2[p_1 - \rho(1-p_1)] < p_0 + \rho(1-p_0) - p_1 + \rho(1-p_1) = q^{0G} - q^{1B} < 0$, we have $W - z^{LH} < 0$, which is a contradiction. So there is no solution in this case.

(c) If $p_1 + p_0 < 1$, then $F = A(1-q^{1B}) + C(1-p_0-p_1) + J > 0$, which implies $z^{LH} = 0$. Then, by (B.9) we have:

$$W - 0 \leq z^{HL} - 0.$$

since $W - z^{HL} \geq 0$, it must be that $W = z^{HL}$. However, using these values in [IC$_3$] leads to a contradiction. So there is still no solution in this case.

vi. $A > 0$, $B > 0$ and $C = 0$. These imply that

$$\begin{cases} q^{1B}(z^{HL} - z^{HH}) + (1 - q^{1B})(z^{LL} - z^{LH}) = \dfrac{c}{p_1 - p_0} \\ p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) = \dfrac{c}{p_1 - p_0}. \end{cases}$$

Since $q^{1B} > p_0$, by (B.9) we obtain:

$$p_0(z^{HL} - z^{HH}) + (1 - p_0)(z^{LL} - z^{LH}) < q^{1B}(z^{HL} - z^{HH}) + (1 - q^{1B})(z^{LL} - z^{LH}) = \frac{c}{p_1 - p_0}.$$

Therefore,

$$p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) + p_0(z^{HL} - z^{HH}) + (1 - p_0)(z^{LL} - z^{LH}) < \frac{2c}{p_1 - p_0},$$

which contradicts [IC$_4$]. So there is no solution in this case.

vii. $A = 0$, $B > 0$ and $C > 0$. These imply that

$$\begin{cases} p_1(z^{LH} - z^{HH}) + (1 - p_1)(z^{LL} - z^{HL}) = \dfrac{c}{p_1 - p_0} \\ p_0(z^{HL} - z^{HH}) + (1 - p_0)(z^{LL} - z^{LH}) = \dfrac{c}{p_1 - p_0}. \end{cases} \tag{B.13}$$

Further, we have $D = Bp_1 + C(p_1 + p_0)$. Again we need to discuss different values of $p_1 + p_0$.

(a) If $p_1 + p_0 = 1$, then $E = B(1 - p_1) + I > 0$ and $J = Bp_1 + F > 0$, which imply $z^{HL} = 0$ and $W = z^{LH}$. Substituting the values into [IC$_2$] yields a contradiction, so that there is no solution in this case.

(b) If $p_1 + p_0 > 1$, then $J = Bp_1 + C(p_1 + p_0 - 1) + F > 0$, which implies $W = z^{LH}$. Using the value in (B.13) we obtain:

$$W - z^{HL} = -\frac{c}{p_0} < 0,$$

which is a contradiction. So there is no solution in this case.

(c) If $p_1 + p_0 < 1$, then $E = B(1 - p_1) + C(1 - p_1 - p_0) + I > 0$, which implies $z^{HL} = 0$. Then, by (B.9) we have:

$$W - z^{LH} \leq 0 - 0.$$

since $W - z^{LH} \geq 0$, it must be that $W = z^{LH}$. However, using these values in [IC$_2$] leads to a contradiction. So there is still no solution in this case.

viii. $A > 0$, $B > 0$ and $C > 0$. The same argument as in case iv applies here and there is no solution in this case.

Summarizing the above analysis, we have the solution for our problem as follows:

(i) $p_1 + p_0 > 1$	(ii) $p_1 + p_0 = 1$	(iii) $p_1 + q_{1G} \leq 1$
$W = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$W = \frac{2c}{p_1-p_0}$	$W = \frac{1+\rho(1-p_1)}{1-p_1}\frac{c}{p_1-p_0}$
$z^{HH} = 0$	$z^{HH} = 0$	$z^{HH} = 0$
$z^{HL} = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$z^{HL} = \frac{c}{p_1-p_0}$	$z^{HL} = \rho\frac{c}{p_1-p_0}$
$z^{LH} = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$z^{LH} = \frac{c}{p_1-p_0}$	$z^{LH} = 0$
$z^{LL} = \frac{2c}{(p_1+p_0)(p_1-p_0)}$	$z^{LL} = \frac{2c}{p_1-p_0}$	$z^{LL} = \frac{1+\rho(1-p_1)}{1-p_1}\frac{c}{p_1-p_0}$

Q.E.D.

Proof of Proposition 7 & 8 follows from Proposition 7 and Table 2.2.

APPENDIX C

Proofs for Chapter 3

PROOF OF PROPOSITION 9. Given budget equation (3.1):

$$W = r_1^H + r_2^H = r_1^L + r_2^L.$$

Since effort implies a higher probability of achieving signal σ_H, to incentivize agent 1 to exert effort, we must have $r_1^L < r_1^H$. Then by budget equation (3.1), $r_2^L > r_2^H$, which will encourage agent 2 to exert zero effort since shirking implies a higher chance of low signal and thus higher reward (r_2^L). Therefore, with budget balanced as in (3.1), it is impossible to induce both agents to exert effort. **Q.E.D.**

PROOF OF PROPOSITION 10. Given the simultaneous move efforts game in Fig. 3.1, for $(1,1)$ to be a Nash equilibrium the following conditions must be satisfied:

$$\begin{array}{lrcl} & p_2 r_1^H + (1-p_2) r_1^L - c & \geq & p_1 r_1^H + (1-p_1) r_1^L, \\ \text{and} & p_2 r_2^H + (1-p_2) r_2^L - c & \geq & p_1 r_2^H + (1-p_1) r_2^L, \end{array}$$

which can be simplified to:

$$\begin{array}{lrcl} & r_1^H - r_1^L & \geq & \dfrac{c}{p_2 - p_1}, \\ \text{and} & r_2^H - r_2^L & \geq & \dfrac{c}{p_2 - p_1}. \end{array}$$

For $(1, 1)$ to be a unique Nash equilibrium, we need to further impose a dominant strategy condition for one of the agents, say agent 1, to exert effort:

$$p_1 r_1^H + (1 - p_1) r_1^L - c \geq p_0 r_1^H + (1 - p_0) r_1^L,$$

which can be simplified to:

$$r_1^H - r_1^L \geq \frac{c}{p_1 - p_0}.$$

By complementary effort assumption (see [A2]), the above inequalities lead to the following incentive compatibility conditions:

$$\begin{aligned} &\textbf{[IC}_1\textbf{]} \qquad r_1^H - r_1^L \geq \frac{c}{p_1 - p_0} \\ &\textbf{[IC}_2\textbf{]} \qquad r_2^H - r_2^L \geq \frac{c}{p_2 - p_1}. \end{aligned}$$

For each agent to participate, we require the expected payoffs in equilibrium be greater than or equal to the opportunity cost of labor (normalized to zero) for both, that is:

$$\begin{aligned} &\textbf{[PC}_1\textbf{]} \qquad p_2 r_1^H + (1 - p_2) r_1^L - c \geq 0 \\ &\textbf{[PC}_2\textbf{]} \qquad p_2 r_2^H + (1 - p_2) r_2^L - c \geq 0. \end{aligned}$$

Finally, the rewards and money burning must be non-negative:

$$\textbf{[NCs]} \qquad r_i^H \geq 0, \quad r_i^L \geq 0, \quad z^H, z^L \geq 0.$$

It is easy to verify that given $r_i^L \geq 0$, [IC$_1$] and [IC$_2$] imply that $r_1^H \geq 0$ and $r_2^H \geq 0$. On the other hand, since $z^H = 0$ and $W = r_1^H + r_2^H + z^H = r_1^L + r_2^L + z^L$, [IC] conditions also imply that:

$$z^L = r_1^H + r_2^H - r_1^L - r_2^L > 0.$$

Further,

$$\begin{aligned}
\text{LHS of [PC}_1\text{]} &= r_1^L + p_2(r_1^H - r_1^L) - c \\
&\geq r_1^L + p_2 \frac{c}{p_1 - p_0} - c && \text{(by [IC}_1\text{])} \\
&\geq \frac{p_2 - p_1 + p_0}{p_1 - p_0} c && \text{(by [NCs])} \\
&\geq 0, && \text{(by [A1])}
\end{aligned}$$

which shows that [PC$_1$] is implied by [IC$_1$] and [NCs] conditions. The same argument applies to [PC$_2$]. Therefore, the principal's problem can be simplified to:

$$\begin{aligned}
&\min_{r_i^H, r_i^L} && r_1^H + r_2^H && (*) \\
&s.t. && r_1^H - r_1^L \geq \frac{c}{p_1 - p_0}, \quad r_2^H - r_2^L \geq \frac{c}{p_2 - p_1} \\
& && r_1^L \geq 0, \quad r_2^L \geq 0.
\end{aligned}$$

Write the corresponding Lagrangian:

$$L = -r_1^H - r_2^H + A\left[r_1^H - r_1^L - \frac{c}{p_1 - p_0}\right] + B\left[r_2^H - r_2^L - \frac{c}{p_2 - p_1}\right] + Cr_1^L + Dr_2^L.$$

The first-order conditions are:

$$\begin{aligned}
\frac{\partial L}{\partial r_1^H} &= -1 + A = 0, & \frac{\partial L}{\partial r_1^L} &= -A + C = 0 \\
\frac{\partial L}{\partial r_2^H} &= -1 + B = 0, & \frac{\partial L}{\partial r_2^L} &= -B + D = 0,
\end{aligned}$$

which imply $A = C = B = D = 1$. Therefore, by the complementary slackness condition, all the constraints in Problem ($*$) are binding:

$$\begin{aligned} r_1^H - r_1^L &= \frac{c}{p_1 - p_0}\,, \qquad & r_1^L &= 0, \\ r_2^H - r_2^L &= \frac{c}{p_2 - p_1}\,, \qquad & r_2^L &= 0. \end{aligned}$$

Thus, the solution to the principal's problem can be summarized as follows:

$$\begin{cases} W^{MB} = \dfrac{c}{p_1 - p_0} + \dfrac{c}{p_2 - p_1} \\ r_1^H = \dfrac{c}{p_1 - p_0}\,, \quad r_2^H = \dfrac{c}{p_2 - p_1}\,, \quad z^H = 0 \\ r_1^L = 0\,, \quad r_2^L = 0\,, \quad z^L = \dfrac{c}{p_1 - p_0} + \dfrac{c}{p_2 - p_1}. \end{cases}$$

Q.E.D.

PROOF OF PROPOSITION 11. The principal's objective is to implement $(1, 1, 0)$ at minimal costs. From Lemma 4, no additional incentive is needed for agent 3.

For $(1, 1, 0)$ to be a unique Nash equilibrium, the incentive compatibility conditions for agents 1 and 2 are the same as in the money burning case:

$$\begin{aligned} &\textbf{[IC}_1\textbf{]} \qquad r_1^H - r_1^L \geq \frac{c}{p_1 - p_0} \\ &\textbf{[IC}_2\textbf{]} \qquad r_2^H - r_2^L \geq \frac{c}{p_2 - p_1}. \end{aligned}$$

Given the non-negativity of rewards, it can be shown that agents 1 and 2's participation constraints are implied by the [ICs], using a similar argument as in the proof of the optimal money burning mechanism. Agent 3's participation constraint in equilibrium is also guaranteed by the non-negativity of r_3^H and r_3^L, since he does not exert any effort.

Since $W = r_1^H + r_2^H + r_3^H = r_1^L + r_2^L + r_3^L$, r_3^L is determined by the other choice variables:

$$r_3^L = r_1^H + r_2^H + r_3^H - r_1^L - r_2^L \geq 0.$$

Again, non-negativity of r_i^H is implied by [IC] conditions.

Now the principal's problem can be written as:

$$\begin{aligned}
\min_{r_1^H, r_1^L, r_2^H, r_2^L, r_3^H} \quad & r_1^H + r_2^H + r_3^H \\
s.t. \quad & r_1^H - r_1^L \geq \frac{c}{p_1 - p_0}\,, \quad r_2^H - r_2^L \geq \frac{c}{p_2 - p_1} \\
& r_1^L \geq 0\,, \quad r_2^L \geq 0\,, \quad r_3^H \geq 0 \\
& r_1^H + r_2^H + r_3^H - r_1^L - r_2^L \geq 0.
\end{aligned}$$

Using the standard Kuhn-Tucker method to solve the principal's problem yields:

$$\begin{cases}
W^U = \dfrac{c}{p_1 - p_0} + \dfrac{c}{p_2 - p_1} \\
r_1^H = \dfrac{c}{p_1 - p_0}\,, \quad r_2^H = \dfrac{c}{p_2 - p_1}\,, \quad r_3^H = 0 \\
r_1^L = 0\,, \quad r_2^L = 0\,, \quad r_3^L = \dfrac{c}{p_1 - p_0} + \dfrac{c}{p_2 - p_1}.
\end{cases}$$

Q.E.D.

PROOF OF PROPOSITION 12. In the text, we have already shown that sabotage-proofness will fail for *unique Nash implementation*. To show a similar result with *weak Nash* or *dominant strategy implementation*, we only need to change the incentive compatibility conditions for agents 1 and 2 to correspond to the respective implementation type. Then by exactly the same argument, we obtain the non-existence result. **Q.E.D.**

Bibliography

[1] Andreoni, James (1991). Reasonable doubt and the optimal magnitude of fines: Should the penalty fit the crime? *RAND Journal of Economics*, 22, 385-395.

[2] Aoyagi, Masaki (2010). Information feedback in dynamic tournament. *Games and Economic Behavior*, 70, 242-260.

[3] Bag, Parimal K. and Neng Qian (2013). “Revisiting money burning in performance evaluation.” Paper presented at the 7th Annual Japan-Taiwan Conference in Contract Theory at Academia Sinica, Taipei.

[4] Baker, George, Robert Gibbons, and Kevin J. Murphy (1994). Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics*, 109, 1125-1156.

[5] Chan, Jimmy and Bingyong Zheng (2011). Rewarding improvements: Optimal dynamic contracts with subjective evaluation. *RAND Journal of Economics*, 42, 758-775.

[6] Eswaran, Mukesh and Kotwal, Ashok (1984). The moral hazard of budget-breaking. *The RAND Journal of Economics*, 15, 578-581.

[7] Fang, Hanming and Giusepe Moscarini (2005). Morale hazard. *Journal of Monetray Economics*, 52, 749-777.

[8] Fuchs, William (2007). Contracting with repeated moral hazard and private evaluations. *American Economic Review*, 97, 1432-1448.

[9] Goltsman, Maria and Arijit Mukherjee (2011). Interim performance feedback in multistage tournaments: The optimality of partial disclosure. *Journal of Labor Economics*, 29, 229-265.

[10] Harris, Milton and Artur Raviv. (1979). Optimal incentive contracts with imperfect information. *Journal of Economic Theory*, 20, 231-259.

[11] Holmstrom, Bengt. (1979). Moral hazard and observability. *Bell Journal of Economics*, 10, 74-91.

[12] Kahn, Lawrence and Peter Sherer (1990). Contingent pay and managerial performance. *Industrial Labor Relations Review*, 43, 107-121.

[13] Konrad, Kai (2000). Sabotage in rent-seeking contests. *Journal of Law, Economics and Organization*, 16, 155-165.

[14] Lazear, Edward P. (1989). Pay equality and industrial politics. *Journal of Political Economy*, 97, 561-580.

[15] Levin, Jonathan (2003). Relational incentive contracts. *American Economic Review*, 93, 835-857.

[16] Lizzeri, Alessandro, Margaret Meyer, and Nicola Persico (2002). “The incentive effects of interim performance evaluations.” Unpublished manuscript (CARESS Working Paper #02-09).

[17] MacLeod, W. Bentley (2003). Optimal contracting with subjective evaluation. *American Economic Review*, 93, 216-240.

[18] Milgrom, Paul R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12, 380-391.

[19] Murphy, Kevin J. and Paul Oyer (2003). “Discretion in executive incentive contracts.” available at http://faculty-gsb.stanford.edu/oyer/wp/disc.pdf.

[20] Neal, Derek (2011). The design of performance pay in education. *Handbook of Economics of Education*, 4, 495-548 (editors: Eric A Hanushek, Stephen J. Machin, Ludger Woessmann), North-Holland publication.

[21] Neal, Derek and Gadi Barlevy (2012). Pay for percentile. *American Economic Review*, 102, 1805-1831.

[22] Prendergast, Canice (1993). A theory of “Yes Men”. *American Economic Review*, 83, 757-770.

[23] Prendergast, Canice (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37, 7-63.

[24] Prendergast, Canice and Robert Topel (1996). Favoritism in organizations. *Journal of Political Economy*, 104, 958-978.

[25] Polinsky, Mitchell A. and Steve Shavell (1979). The optimal trade-off between the probability and magnitude of fines. *American Economic Review*, 69, 880-891.

[26] Rajan, Madhav V. and Stefan Reichelstein (2006). Subjective performance indicators and discretionary bonus pools. *Journal of Accounting Research*, 44, 585-618.

[27] Rajan, Madhav V. and Stefan Reichelstein (2009). Objective versus subjective indicators of managerial performance. *The Accounting Review*, 84, 209-237.

[28] Rasmusen, Eric B. (1987). Moral hazard in risk-averse teams. *The RAND Journal of Economics*, 18, 428-435.

[29] Sebastian, Goerg J., Kube Sebastian, and Ro'i Zultan. (2010). Treating equals unequally: Incentives in teams, workers' motivation, and production technology. *Journal of Labor Economics*, 28, 747-772.

[30] Winter, E. (2004), Incentives and discrimination. American Economic Review, 94, 764-773.

[31] Zabojnik, Jan (2011). Subjective evaluations with performance feedback. Forthcoming in *RAND Journal of Economics* (Queen's Economics Department Working Paper No. 1283).