Optimality of Quota Contract

Pak-Hung Au, Keiichi Kawai†

October 10, 2014

Abstract

A quota contract – characterized by a target, and discrete and sizable reward for achieving it – is susceptible to gaming by the agent, for example, manipulating the timing of closing a deal and reporting earnings. Because of its obvious drawback, the widespread use of quota contracts for salespeople and executives has puzzled economists. In this paper, we show that using a quota contract can be optimal when the principal has a contract commitment problem, and the optimality of its use arises precisely from the agent’s incentive-system gaming.

Introduction

A quota-based contract – a form of incentive contract characterized by a production or a profit target as well as a discrete and sizable reward for achieving it – is a prevalent form of incentive contract, especially for salespeople and executives. The widespread use of quota-based contracts, however, has puzzled economists because the nonlinearity of quota-based contract is known to be susceptible to “system gaming”, i.e., agents manipulating the timing of business to increase his overall compensation.

Empirical and anecdotal evidence abounds for system gaming activities, and the associated costs it imposes on the firms. For example, by looking into executives’ accrual decision, Healy (1985) finds that executives do not report

1

According to Joseph and Kalwani (1998), quota-based compensation is one of the most consistent features in the sales industry. Specifically, Oyer (2000) notes that over 80% of salespeople are compensated according to a quota scheme.
all earnings if the firm’s performance in the current fiscal year is so bad that they are unlikely to reach the earning target. Also, Oyer (1998) points out that salespeople often have the ability to influence when deals are closed and sales are generated.\(^2\) He finds evidence that under quota compensation schemes, salespeople play “timing games”: they “pull in” business from the next fiscal period if they are about to meet the quota this period; or “push out” business to the next fiscal period if they fall too far behind this period’s quota. The timing manipulation by agents results in a spike in output at the end of each period, and a dip at the beginning of each period, causing the firm undesirable output fluctuation. In a similar vein, Larkin (2013) investigates the sales pattern of a large computer software firm and finds that gaming is widespread and costly to the firm. According to his estimation, price discounts offered to customers due to agents’ gaming cost the firm 6-8% of total revenue.

These empirical evidence against quota-based contracts, however, overlook a hidden benefit of incentive-system gaming to the firms. In this paper, we show that the firm can save agency cost by deliberately inducing agents to game the system. Our theory thus sheds light on the widespread use of quota-based contracts in the industry.

We analyze a dynamic moral hazard model in which a risk-neutral (female) principal hires a risk-neutral (male) agent to engage in production for two fiscal periods. In every period, the agent receives a nonnegative wage. In period one, the agent engages in two rounds of production. The principal, however, can neither directly observe the agent’s actual production results nor distinguish outputs produced in different periods. She can only observe and verify outputs that are voluntarily turned in by the agent. The agent therefore has rooms to privately store outputs produced in period one and turn in them in period two where the agent engages in another round of production.\(^3\)

The principal can easily discourage the agent from engaging in such timing manipulation by rewarding him according to a linear contract: a constant and sufficiently large bonus is paid for every unit of reported output. On the other hand, should the principal use a quota-based contract, the agent who marginally misses the quota has every incentive to “push out” his output to period two, which imposes a direct time-discounting cost the principal. At first sight, the principal is always hurt by agents’ system gaming, and a quota-based contract is dominated by a more linear contract.

To understand our finding, note that in a dynamic moral hazard setting, the principal can motivate the agent via both current incentives (immediate payments for good performance) and future incentives (threat of lower future incentive payment or even termination of the relationship following poor performance).\(^4\) Although using future incentives is costly to the principal as

\(^2\)For example, a salesperson can control when to close a deal by deciding when to offer price discounts to customers. If the salesperson does not have full authority over price setting, he/she can persuade managers in charge that price discounts are necessary.

\(^3\)This is a natural assumption for salespersons and division managers. A salesperson often works exclusively with clients, making it difficult to for the principal to monitor the agent’s progress. A division manager often has some control over information flow to higher management.
valuable agent effort is forgone, the threat of punishment for poor performance help motivate the agent. Therefore, future incentive is useful for saving agency payments, and the optimal long-term contract typically combines the use of both sources of incentives.

Our theory is built on the observation that a principal lacking power to contract commitment may not be able to use future incentives effectively. At the interim stage following bad performance, the principal has every incentive to reward the agent for good future performance (which results in high subsequent effort) rather than punishing his poor past performance (which results in low subsequent effort). In other words, the contractual punishment on the agent may not be credible as both the principal and agent are better off if they agree to improve the efficiency of the existing contract. To capture the principal’s temptation to improve the existing contract, we consider a contracting environment in which the principal and the agent can enter into a long-term contract, but at the beginning of each fiscal period, they can renegotiate to replace the existing contract. For simplicity, the renegotiation game is modelled as the principal making a take-it-or-leave-it offer to the agent.\footnote{That is, if the new offer is accepted by the agent, it replaces the existing contract; if rejected, the existing contract remains in effect. See for example, Fudenberg and Tirole (1990) and Hermalin and Katz (1991).}

Below, we provide the intuition of why quota-based contract helps making punishment credible. Consider the following long-term contract: (i) a first-period bonus is paid if and only if the quota is met; (ii) if the first-period quota is NOT met, the second-period quota is so high that it can be met only if the agent carries over sufficiently many outputs from the first period. Under such contract, if an agent fails to meet the first-period quota, he naturally carries his outputs over to the second period. Moreover, if he fails by a large margin, he gives up working in the second period.

Now consider the contract renegotiation stage following a failure to meet the first-period quota. The principal can benefit from improving the contract for the agent who failed first-period quota by a large margin. The difficulty for the principal is that she does not know the “type” of the agent she is facing. An agent who has carried over a lot of outputs can pretend to have little outputs and collect the bonus without putting in any further effort. As a result, such contractual improvement may entail extra information rent to solicit truth-telling by the agent. The principal therefore is deterred from improving the contract if the extra (expected) information rent exceeds the benefit of doing so.

In sum, under the possibility of contract renegotiation, a quota-based contract helps the principal to recover contract-commitment ability. This is possible because it induces the agent to game the incentive system, thus creating an asymmetric information problem for the principal, which raises her implicit cost of contract renegotiation. As a result of the improvement in commitment power, the principal now can enjoy the savings in overall agency cost in comparison to any contract that does not induce any system gaming, including the linear contract. Our theory thus provides a unified explanation for the use of
quota-based contracts, and the associated gaming activities by the agent. In sharp contrast to the existing literature which views agent’s system gaming as a dysfunctional response to quota-based contracts (see the survey in Prendergast (1999)), our theory proposes that agent’s system gaming can be beneficial to the principal, and she may deliberately design a contract and/or work environment in order to encourage such activities.

We provide sufficient conditions under which the optimal renegotiation-proof contract takes a quota form. Loosely speaking, these conditions ensure that a quota contract (i) is renegotiation-proof; (ii) implement the effort profile in the optimal full-commitment contract; and (iii) involves a small cost due to agent’s gaming.

Our result suggests that the principal can be strictly better off by having less information about the agent’s production results. This finding is in sharp contrast to Holmstrom (1979)'s Informativeness Principle, which states that the principal always benefits from having more informative contractible signals about the agent’s effort choice. The reason for this difference in conclusions is that Holmstrom (1979) considers contracting under full commitment, while we assume the principal lacks commitment. In our model, by choosing to learn less, the principal regains the power to commit to the use of future incentives, and it in turn helps her lower the overall agency costs and improve profit.

The idea that the principal can benefit from having less information about the agent’s production is reminiscent of Cremer (1995). He analyzes a situation where a principal faces an agent whose output depends on both his exogenous type and effort. He then shows that by refraining from learning the agent’s type at the interim stage, the principal can commit to punish the agent severely following bad performance, thus lower the agency cost. In contrast, in our pure moral hazard setting in which the agent has no exogenous type, the principal endogenously creates agent types and hence increases her commitment ability by using a quota-based contract and inducing the agent to game the system.

Several other papers have proposed explanations for the use of quota contracts. Oyer (2000) shows the optimality of quota contract in a static moral hazard setting. While his finding is sensitive to the assumption that the agent is risk-neutral as shown by Jewitt et al. (2008), our results can be shown to be robust to agent’s risk aversion. Herweg et al. (2010) shows that quota contract is optimal if the agent is expectation-based loss averse. Larkin and Leider (2012) suggest that convex payment scheme can attract over-confident workers, thus help the firm save agency costs. Our theory, in contrast, is based on standard preference and full rationality of the agent. Moreover, all existing work mentioned above focus on static settings. On the other hand, by considering a dynamic model, we identify a novel source of benefit of the quota contract: it induces the agent to game the system, which in turn mitigates a commitment problem faced by the principal.

Finally, our dynamic moral hazard model features risk-neutral principal and agent, and the source of agency conflict is limited liability of the agent. Similar setup has been explored in, for example, Ohlendorf and Schmitz (2012) and Fong and Li (2012). Unlike our model, these papers do not allow gaming
activities by the agent: all outputs are immediately available the principal once they are produced.

Model

The game consists of two periods, $t = 1, 2$. There are two risk-neutral players: a (female) principal, and a (male) agent.

The first period, $t = 1$, consists of two rounds of production, and the second period, $t = 2$, consists of one round of production. In each of these stages, the agent chooses $e \in \{0, 1\}$. The choice of $e$ is unobservable by the principal, but the cost of effort in period $t$, $c_t(e) = c_t e$, is commonly known. The (marginal) cost of effort is increasing over time, that is $c_2 > c_1$.

The agent’s effort level stochastically affects the output $y \in \{0, 1\}$. Let $p_e \in (0, 1)$ denote the probability of outcome $y = 1$ given effort $e$, where $p_0 \equiv p_1 - \Delta$ for some $\Delta \in (0, p_1)$. We refer to $e = 1$ and $e = 0$ as high and low efforts, respectively. The outcome of each round is independent. We use $y_t$ to denote the total output in period $t$. Note that $y_1 \in \{0, 1, 2\}$, and $y_2 \in \{0, 1\}$. We normalize the value of one unit of output to the principal to one, and to the agent to zero.

At the beginning of the game, the principal chooses one of two possible monitoring technologies, which will call the “agent self-report scheme” and the “direct monitoring.” If the principal chooses the agent self-report scheme, she observes neither the agent’s total output at the end of each period, nor the round at which the output is made within $t = 1$. At the end of each period, however, the agent can report the total output to the principal. The reported output is verifiable and hence is contractible. That is, in $t = 1$, the agent can report $r_1 \in \mathbb{Z}_+$ if and only if $r_1 \leq y_1$. When agent reports $r_1 < y_1$, he can hide and store $y_1 - r_1$ to $t = 2$.

If the principal chooses the direct monitoring technology, she can perfectly observe and verify the agent’s total output at the end of each period. However, the exact round in $t = 1$ at which an output is made remains neither observable nor verifiable. Therefore, the timing of the output within $t = 1$ is not contractible. When the direct monitoring technology is adopted, we can, without loss of generality, assume that the agent’s report $r_t$ on the output $y_t$ in period $t$ satisfies $r_t = y_t$.

Once adopted, the choice of monitoring technology becomes common knowledge. Also it is commonly known that the technology is fixed for the rest of the game.

After the choice of the monitoring technology is made, and becomes common knowledge, the principal offers a long-term contract to the agent. The contract specifies the first-period bonus $b_1(r_1)$ that depends on the reported output in $t = 1$, and the second-period bonus $b_2(r_2; r_1)$ that depends both on the first-period and second-period reports.

---

Salesforce, for example, is likely to need to exert more effort to generate later sales than earlier sales because he is likely to deplete “easier” sales.
Following the report \( r_1 \) in \( t = 1 \), the principal may renegotiate with the agent by making a take-it-or-leave-it offer. The offer consists of a menu of payment plans for the second period. If the agent selects a new plan \( b_2 (r_2) \) from the menu, the new plan replaces the existing contract. If he rejects all plans from the menu, the original contract remains in effect. Without loss of generality, we assume that the menu of contracts includes the original contract, and the agent selects a plan from the menu.

The timing of the game is as follows:

1. The principal chooses the monitoring technology.
2. The principal offers a contract \( \{b_1 (r_1), b_2 (r_2; r_1)\} \), and the agent decides whether to accept the contract or not.
3. The agent chooses \( e \) for the first round of \( t = 1 \).
4. The output for the first round of \( t = 1 \) realizes.
5. The agent chooses \( e \) for the second round of \( t = 1 \).
6. The output for the second round of \( t = 1 \) realizes.
7. The agent makes the report \( r_1 \leq y_1 \), and \( b_1 (r_1) \) will be paid to him.
8. The principal offers a menu of new contract.
9. The agent selects a plan from the menu.
10. The agent chooses \( e \) for \( t = 2 \).
11. The output realizes for \( t = 2 \) realizes.
12. The agent makes the report \( r_2 \), and will be paid according to the payment plan he has chosen at step 9.

Assumption 1

1. In \( t = 1 \), the principal discounts her payoff in \( t = 2 \) by the discount factor \( \delta \in (0, 1) \), while the agent discounts his payoff in \( t = 2 \) by the discount factor \( \rho \in (0, \delta) \).
2. The agent is protected by limited liability. Specifically, the transfers from the principal to the agent in any period is nonnegative.

Remark 1

The first assumption ensures that it is never optimal for the principal to delay the payment. The second assumption ensures that principal cannot sell the production technology to the agent. It also implies that agent cannot borrow money from the principal to take advantage of the difference in discount factors.

An important class of contract is a linear contract.

Definition 1

We say a contract is linear if \( b_1 (r_1) = \alpha_1 \times r_1 \) and \( b_2 (r_2; r_1) = \alpha_{2,r_1} \times r_2 \) for some positive numbers \( \alpha_1 \) and \( \alpha_{2,r_1} \).

\(^6\)See DeMarzo and Fishman (2007), for example.
We are interested in situations where a linear contract is optimal if the principal is myopic: she chooses the optimal contract for each period independently. This translates into the following requirements on effort costs.\footnote{Suppose that the game only consists of period one and the principal can observe the output of each round. Then, the principal’s payoffs of (i) inducing effort in both rounds, (ii) payoff of inducing no effort in both rounds, and (iii) inducing effort if and only if the first round is successful, are \(i) 2p_1\left(1 - \frac{\Delta}{p_1}\right), \(ii) 2(p_1 - \Delta), \) and \(iii) p_1^2\left(2 - \frac{\Delta}{p_1}\right) + (2p_1 - \Delta)(1 - p_1).\) Therefore, if \(c_1 < \frac{(1-p_1)^2}{p_1(1-\Delta)},\) the agent finds it optimal to induce effort in both rounds.}

**Assumption 2** The costs of effort satisfy the following conditions: \((i) c_1 < \tilde{c}_1 \equiv \frac{(1-p_1)\lambda^2}{p_1(1-\Delta)}\) and \((ii) c_2 < \tilde{c}_2 \equiv \frac{\lambda^2}{p_1}\).\footnote{Under the optimal linear contract, the agent’s payoffs from gaming the system and not gaming the system are \(p_1\lambda^2\Delta + c_1/\Delta + p(p_1 - \Delta)c_2/\Delta\). Therefore, the condition under which the agent does not game the system under the optimal linear contract boils down to \(c_1/c_2 < p(1-(p_1-\Delta))\).}

With a linear contract, if the slope of the payment scheme does not increase too much over time, the agent does not delay reporting output under the self-reporting scheme. To ensure this, we impose an upper bound on the rate at which the effort cost increases.\footnote{If \(c_1/c_2\) is sufficiently small, then, the future incentives alone can incentive the agent to work in the first period. In contrast, if \(c_1/c_2\) is sufficiently large, the principal finds it optimal to incentivize the agent to work in \(t = 1\) without using future incentives.}

**Assumption 3** The ratio of \(c_1\) and \(c_2\) satisfies the following condition: \(c_1/c_2 \geq \rho(1-(p_1-\Delta))\).

Below, we show that once the dynamic consideration is taken into account, the optimal contract may not be linear. This is because while a linear contract consists of only immediate incentives, the principal may find it useful to also utilize future incentives to lower the overall agency cost. More precisely, the principal can lower the agency cost by giving a low continuation payoff for the agent with a very poor performance in \(t = 1\). Throughout the paper, we limit our attention to the case where the ratio of effort costs \(c_1/c_2\) is neither too small nor too large so that the principal find it useful to also utilize both current and future incentives when she can commit to a long-term contract.\footnote{If \(c_1/c_2\) is sufficiently small, then, the future incentives alone can incentive the agent to work in the first period. In contrast, if \(c_1/c_2\) is sufficiently large, the principal finds it optimal to incentivize the agent to work in \(t = 1\) without using future incentives.}

**Assumption 4** The ratio of \(c_1\) and \(c_2\) satisfies the following condition: \(c_1/c_2 \in (\rho(p_1 - \Delta), \rho)\).
analysis that, by offering a quota-based contract with self-reporting scheme, the principal can deliberately raise the implicit cost of renegotiation, and hence mitigate this commitment problem.

1 Benchmark Case: Full Commitment to Long-term Contract

Suppose the principal can commit to a long-term contract. By Holmstrom’s informativeness principle, the principal is weakly better off with more information, and hence will choose the direct monitoring technology. As a result, there is no room for the agent to game the incentive system.

We start by showing that the use of future incentive involves the trade-off of saving agency cost in $t=1$ and the loss in valuable effort in $t=2$. If the principal wants to induce the effort for the agent in $t=2$, she has to give up an agency rent at least $(p_1 - \Delta) c_2 / \Delta$ by paying $b_2 (1; r_1) = c_2 / \Delta$, irrespective of the performance in $t=1$. Since $c_2 \leq c_2'$, the resulting principal’s payoff $p_1 \left(1 - \frac{c_2}{\Delta}\right)$ is larger than the payoff from not inducing effort $p_1 - \Delta$. Therefore, at the interim stage, the principal always prefers inducing effort in $t=2$.

From an ex-ante point of view, this does not necessarily mean that principal always finds it beneficial to induce effort in $t=2$. In fact, if the principal decides to induce effort following $y_1 > 0$ only, then the agent’s with $y_1 = 0$ loses the chance to earn the agency rent of $(p_1 - \Delta) c_2 / \Delta$ in $t=1$. This in turn reduces the agency cost needed to induce efforts in $t=1$. That is, by punishing the agent in $t=2$ for poor performance in $t=1$, i.e., $y_1 = 0$, then principal can save the agency cost in $t=1$. Throughout the paper, we call a contract that induces no effort following $y_1 = 0$ a termination contract.

Definition 2 A contract is a termination contract if $b_2 (1; 0) = 0$.

Under a termination contract, the losses in valuable effort in $t=2$ and the savings in agency cost relative to the optimal linear contract are, respectively,

\[
C (p_1, c_2) \equiv (1-p_1)^2 \times \delta \times \frac{\left(p_1 \left(1 - \frac{c_2}{\Delta}\right) - (p_1 - \Delta)\right)}{\Pr (y_1=0)} \quad \text{principal’s loss in } t=2 \text{ when } y_1=0.
\]

\[
S (p_1, c_2) \equiv p_1 \left(2-p_1\right) \times \rho \times \frac{\left(\frac{(p_1 - \Delta) c_2}{\Delta}\right)}{\Pr (y_1\neq0)} \quad \text{agent’s rent in } t=2 \text{ when } y_1\neq0.
\]

\[\text{Given bonuses of } b_2 (1; r_1) = b_2 \text{ and } b_2 (0; r_1) = 0, \text{ the agent is willing to put in effort in } t=2 \text{ if and only if } p_1 b_2 - c_2 \geq p_0 b_2. \text{ Hence, } b_2 (1; r_1) = \frac{c_2}{\Delta}, \text{ and the resulting agent’s payoff is.}
\]

$p_1 b_2 (1; r_2) - c_2 = \frac{(p_1 - \Delta) c_2}{\Delta}$.\]
Whether the benefit of saving agency cost in \( t = 1 \) outweighs the loss in valuable effort in \( t = 2 \) in a termination contract depends on the size of \( c_2 \). The larger the value of \( c_2 \), the lower the principal’s benefit from inducing effort in \( t = 2 \), and the lower the losses in valuable effort \( C(p_1, c_2) \). In contrast, from the agent’s perspective, a large value of \( c_2 \) means a large agency rent \((p_1 - \Delta)c_2/\Delta\) in \( t = 2 \). Therefore, it is relatively easy for the principal to induce the agent to work in \( t = 1 \) by threatening to fire the agent if he does not perform well in \( t = 1 \), and the savings in agency cost \( S(p_1, c_2) \) are large.

In the same vein, whether the benefit of saving agency cost in \( t = 1 \) outweighs the loss in valuable effort in \( t = 2 \) in a termination contract also depends on the size of \( p_1 \). If \( p_1 \) is large, then the probability that the agent fails to produce output in both rounds in \( t = 1 \) is sufficiently small, thus the losses in effort in \( t = 2 \) are small. At the same time, from the agent’s perspective, larger \( p_1 \) implies larger rent \((p_1 - \Delta)c_2/\Delta\) in \( t = 2 \). Moreover, under a termination contract, \( \Pr(y_1 \neq 0) \) is sufficiently large for a sufficiently large \( p_1 \). Hence, the savings in agency cost are sufficiently large. As a result, the principal prefers a termination contract over a linear contract. These observations are formalized in the following theorem.

**Theorem 1**

1. Under the optimal linear contract, the bonuses in \( t = 1 \) are \( b_{1,LC}^L(r_1) = r_1c_1/\Delta \).
   Under the optimal termination contract, the bonuses in \( t = 1 \) are \( b_{1,TC}^L(1) = (c_1 - \rho(p_1 - \Delta)c_2)/\Delta \) and \( b_{1,TC}^L(2) = (2c_1 - \rho(p_1 - \Delta)c_2)/\Delta \).
2. For any \( c_1 \) and \( c_2 \), there exists a \( p_{FC}^1(c_2) \) such that (i) the termination contract is an optimal contract for all \( p_1 \geq p_{FC}^1(c_2) \) and (ii) the linear contract is an optimal contract for all \( p_1 < p_{FC}^1(c_2) \).
3. For any \( c_1 \) and \( p_1 \), there exists a \( c_{2,FC}^L(p_1) \) such that (i) the optimal termination contract is an optimal contract for all \( c_2 > c_{2,FC}^L(p_1) \); and (ii) the optimal linear contract is an optimal contract for all \( c_2 \leq c_{2,FC}^L(p_1) \).

**Proof.** In the Appendix. ■

**Remark 2** Note that neither \( p_{1,FC}^1(c_2) \) nor \( c_{2,FC}^L(p_1) \) depends on \( c_1 \). This is because neither the losses from valuable effort nor savings in agency cost of the termination contract relative to the linear contract depend on the absolute size of \( c_1 \).

## 2 Renegotiation-Proof Contract

Now we move to the main analysis in which the principal may renegotiate with the agent at the beginning of \( t = 2 \). It is without loss of generality to focus on renegotiation-proof contracts, that is, long-term contracts that are immune to renegotiation at the interim stage.
2.1 Direct Monitoring

Suppose that the principal has chosen the direct monitoring technology, which allows the principal to fully observe $y_1$, the agent’s total output in $t = 1$.

We first explain that under direct monitoring, a contract is renegotiation-proof if and only if the contract induces effort in $t = 2$, irrespective of the agent’s performance in $t = 1$. To see this, consider a long-term contract that does not induce agent’s effort in $t = 2$ following some realization of $y_1$. The agent’s effort, however, is always valuable to the principal. Thus, the principal and the agent can always agree on a new continuation contract such that (i) it induces effort in $t = 2$; and (ii) it makes both players better off than under the existing continuation contract. That is, a renegotiation-proof contract must induce effort in $t = 2$. Conversely, if a long-term contract always induces effort in $t = 2$, irrespective of the agent’s performance in $t = 1$, then there is no other contract that can make both players better off. Therefore, the contract is necessarily renegotiation-proof.

This simple observation implies that under direct monitoring, any termination contract is not renegotiation-proof. Furthermore, because of the assumption that inducing effort in both rounds of $t = 1$ is optimal in the absence of dynamic consideration, the optimal renegotiation-proof contract must induce effort in all output contingencies. Since the least costly contract that implements this effort profile is the optimal linear contract, we arrive at the following conclusion.

**Theorem 2** If the direct monitoring technology is adopted, the optimal linear contract is an optimal renegotiation-proof contract.

**Remark 3** If a linear contract is optimal under the full commitment assumption, the renegotiation-proofness constraint has no bite, no matter which monitoring technology is adopted. Interesting results thus arise only if linear contract is not optimal under full commitment. We limit our attention to this case throughout the rest of the paper.

2.2 Self-Reporting Scheme

Now suppose that the principal has chosen the agent self-reporting technology. A form of contract we are particularly interested in is quota contract:

**Definition 3** A contract is a quota contract if for some $B$ and $\beta$,

- $b_1(0) = b_1(1) = 0; b_1(2) = B > 0$
- $b_2(2; r_1) \equiv \beta > c_2/\Delta; b_2(1; r_1) = b_2(0; r_1) = 0$ for $r_1 \in \{0, 1\}$
- $b_2(1; 2) = c_2/\Delta; b_2(0; 2) = 0$.

Under a quota contract, the agent is paid in period one if and only if he turns in two units of output. Thus, the “quota” of the first period is two, and the bonus of meeting the quota is $B$. The payment in period two depends on whether first-period quota is met or not. If the agent meets the quota in the
first period, his second-period quota is one and the associated bonus is \( \frac{c_2}{\Delta} \). Otherwise, his period-2 quota would be two and the associated bonus is \( \beta \).

Use of a quota contract has the following implications. By choosing \( \beta \) and \( B \) appropriately (assuming the quota contract is renegotiation-proof), it induces the same effort profile as the termination contract under full commitment. That is, the agent is induced to exert effort in both rounds in \( t = 1 \), and in \( t = 2 \) if and only if \( y_1 = 1, 2 \). However, different from the termination contract, (i) agent with \( y_1 = 1 \) may game the incentive system, i.e., he may carry his output over to \( t = 2 \); and (ii) as a result, after receiving a report \( r_1 = 0 \), the principal is uncertain whether the agent’s first-period output \( y_1 \) is 0 or 1.

The principal may want to renegotiate with the agent at the beginning of \( t = 2 \). In particular, she may want to solicit effort by agent with \( y_1 = 0 \). The agent with \( y_1 = 0 \) is not be paid any bonus, irrespective of his performance in \( t = 2 \), and thus not putting in any effort in \( t = 2 \). The gains from renegotiation arise from the effort by the agent with \( y_1 = 0 \) thus can be written as

\[
\frac{(1 - p_1)^2}{2p_1 (1 - p_1) + (1 - p_1)^2} \times \left( p_1 \left( 1 - \frac{c_2}{\Delta} \right) - (p_1 - \Delta) \right). \quad \text{principal’s loss in } t=2 \text{ when } y_1=0.
\]

Improving the contract for agent with \( y_1 = 0 \), however, involves an indirect cost: extra information rent must be offered to agent with \( y_1 = 1 \), as the agent’s first period output \( y_1 \) is his private information. More precisely, let \( \{b_2^{y_1} (r_2)\}_{y_1=0} \) be the menu of contracts offered at renegotiation. If the principal decides to induce effort by agent with \( y_1 = 0 \), the payment scheme has to satisfy \( b_2^{y_1} (1) = \frac{c_2}{\Delta} \). At the same time, to prevent agent with \( y_1 = 1 \) from shirking in \( t = 2 \), i.e., prevent him from misrepresenting his type as \( y_1 = 0 \), \( b_2^{y_1} (2) \) has to be set sufficiently high. The costs of renegotiation capture this extra information rent for the agent with \( y_1 = 1 \), which can be written as

\[
\frac{2p_1^2 (1 - p_1)}{2p_1 (1 - p_1) + (1 - p_1)^2} \times p_1 \left( \frac{c_2 \rho - c_1}{p_1 \rho \Delta} \right). \quad \text{info rent for } y_1=1.
\]

The quota-contract is renegotiation-proof when the costs of renegotiation outweigh the gains from renegotiation.

**Lemma 1** If there exists an renegotiation-proof quota contract, then the principal set \( \beta^* \) and \( B^* \) as follows:

\[
\beta^* = \frac{\rho \Delta c_2 + c_1}{p_1 \rho \Delta} \quad \text{and} \quad B^* = \begin{cases} 
B^F & \text{if } p_1 \geq \hat{p}_1 \equiv \frac{\rho c_2 \Delta + c_1}{2c_1} \\
B^I & \text{if } p_1 < \hat{p}_1
\end{cases}
\]

\[
\beta^F \equiv \frac{2c_1 - \rho (p_1 - \Delta) c_2}{\Delta} \quad \text{and} \quad B^I \equiv \frac{\rho (\Delta - (p_1 - \Delta) p_1) c_2 + c_1}{p_1 \Delta},
\]

(2)
A renegotiation-proof quota contract exists if and only if
\[
\frac{2p_1^2 (1 - p_1)}{2p_1 (1 - p_1) + (1 - p_1)^2 p_1 \left( \frac{c_2 p - c_1}{p_1 \rho \Delta} \right)} \geq \frac{(1 - p_1)^2}{2p_1 (1 - p_1) + (1 - p_1)^2} (\Delta - p_1 c_2 / \Delta). \tag{3}
\]

Costs of Renegotiation \hspace{1cm} Gains from Renegotiation

Proof. In the Appendix. \qed

Now we see when a quota contract is renegotiation-proof. First, suppose that \( p_1 \) is sufficiently high. Following a zero output report in period 1, \( r_1 = 0 \), the principal believes that it is quite likely that the agent has \( y_1 = 1 \). That is, \( \Pr(y_1 = 1|r_1 = 0) \) is sufficiently large while \( \Pr(y_1 = 1|r_1 = 0) \) is sufficiently small. Therefore, renegotiation is likely to be unprofitable, and thus the quota contract is likely to be renegotiation-proof.

Similarly, a large value of \( c_2 \) makes the quota contract more likely to be renegotiation-proof for two reasons. First, the gains in the effort by agent with \( y_1 = 0 \) is small (that is, gains from renegotiation are small). Second, should renegotiation occur, the bonus for one unit of output in \( t = 2 \), \( c_2 / \Delta \), is large. Therefore, preventing agent with \( y_1 = 1 \) from mimicking the agent with \( y_1 = 0 \) becomes more expensive (costs of renegotiation are large). These observations are summarized in the following theorem.

**Theorem 3** Consider the quota contract with \( \beta = \beta^* \) and \( B = B^* (\beta^*) \).

1. For arbitrary \( c_1 \) and \( c_2 \), there exists a \( p_1^{RP} (c_1, c_2) < 1 \) such that a renegotiation-proof quota contract exists if and only if \( p_1 \geq p_1^{RP} (c_1, c_2) \).

2. For arbitrary \( p_1 \) and \( c_1 \), there exists \( c_2^{RP} (p_1, c_1) \) such that a renegotiation-proof quota contract exists if and only if \( c_2 \geq c_2^{RP} (p_1, c_1) \).

Proof. In the Appendix. \qed

**Remark 4** Note that by assumption, \( c_2 < \min \left\{ \frac{\bar{c}_2}{\rho (1 - (p_1 - \Delta))}, \frac{\bar{c}_1}{\rho (p_1 - \Delta)} \right\} \). There exists an \( c_1^{RP} (p_1) \) such that \( c_2^{RP} (p_1, c_1) < \min \left\{ \bar{c}_2, \frac{\bar{c}_1}{\rho (1 - (p_1 - \Delta))}, \frac{\bar{c}_1}{\rho (p_1 - \Delta)} \right\} \) if and only if \( c_1 \geq c_1^{RP} (p_1) \).\(^{11}\)

### 2.3 Quota Contract vs Linear Contract

Now we compare the performance of quota contract and linear contract. Recall the optimal linear contract gives the principal the highest payoff under direct

\(^{11}\)First, \( c_2^{RP} (p_1, c_1) \) if and only if \( c_1 \leq \rho \Delta / p_1 \), which always follows from \( c_1 \leq c_1 \). Next, \( c_2^{RP} (p_1, c_1) \) if and only if \( c_1 > p_1^{1+\Delta p_1}/p_1 (1+3p_1+2\Delta) \). Lastly, \( c_2^{RP} (p_1, c_1) \) if and only if \( c_1 > p_1^{1+\Delta p_1}/p_1 (1+3p_1+2\Delta) \). Therefore, for \( c_2^{RP} (p_1) = \frac{\bar{c}_2 p_1}{(1-3p_1+2\Delta)} \) \( \times \max \left\{ \frac{p_1+\Delta}{(1+3p_1+2\Delta)}, \frac{1}{(p_1-\Delta)} \right\} \), \( c_2^{RP} (p_1, c_1) \) if and only if \( c_1 \geq c_1^{RP} (p_1) \).
monitoring. Thus, if the quota contract is renegotiation-proof under agent’s self-report scheme and is able to generate a higher principal payoff, then the optimal monitoring technology is to rely on agent’s self-reporting. While this is impossible under full-commitment contracting according to Holmstrom’s Informativeness Principle, we show that this can happen under the possibility of renegotiation.

Suppose \( S(p_1, c_2) - C(p_1, c_2) > 0 \) so that if contract commitment is perfect, the termination contract is optimal. If renegotiation is possible, the principal can still implement the same effort profile using a quota contract, albeit at a higher cost. The \textit{losses from the lack of commitment}, \( L(p_1, c_1, c_2) \), is defined as the difference in principal’s payoffs under the termination contract and optimal quota contract. We can express the condition for quota contract outperforming linear contract as follows:

\[
\frac{S(p_1, c_2)}{C(p_1, c_2)} - 1 > \frac{L(p_1, c_1, c_2)}{C(p_1, c_2)},
\]

where

\[
S(p_1, c_2) - C(p_1, c_2) > L(p_1, c_1, c_2),
\]

\( S \) Saving in Agency Cost  \hspace{1cm} \( C \) Loss in Valuable Effort  \hspace{1cm} \( L \) Losses from Lack of Commitment

The left-hand side of the inequality is the margin by which the payoff of the termination contract exceeds the linear contract. The right-hand side is the margin by which the payoff of the optimal quota contract falls short of the termination contract.

In order to understand when the inequality (4) holds, we can decompose the losses from lack of commitment \( L(p_1, c_1, c_2) \) into three parts: (i) the cost of inducing truthful report from the agent with \( y_1 = 2 \); (ii) the cost associated with the system gaming by the agent with \( y_1 = 1 \); and (iii) the cost associated with delaying the payment for the agent. More specifically,

\[
L(p_1, c_1, c_2) = L_1(p_1, c_1, c_2) + L_2(p_1, c_1, c_2) + L_3(p_1, c_1, c_2),
\]

where

\[
L_1(p_1, c_1, c_2) \equiv p_1 \max \left\{ \frac{\rho \Delta c_2 + (1 - 2p_1) c_1}{p_1 \Delta}, 0 \right\},
\]

(i) cost of inducing truthful report in \( t = 1 \)

\[
L_2(p_1, c_1, c_2) \equiv 2p_1 (1 - p_1) (1 - \delta), \quad \text{and}
\]

(ii) cost of system gaming

\[
L_3(p_1, c_1, c_2) \equiv 2p_1 (1 - p_1) \left( \frac{(\delta - \rho) c_1 - \rho (p_1 - \Delta) c_2}{\rho \Delta} \right),
\]

(iii) cost of delaying payment

The cost of inducing truthful report, \( L_1(p_1, c_1, c_2) \), arises from preventing the agent with \( y_1 = 2 \) to withhold the produced outputs to \( t = 2 \). It is zero if \( p_1 \) is sufficiently large, or \( c_1 \) is not sufficiently small relative to \( c_2 \). First, if \( p_1 \) is sufficiently large, then the agent with \( y_1 = 2 \) is likely to receive an additional
bonus by exerting effort in $t = 2$, and hence does not have incentive to withhold the output in $t = 1$.\footnote{More specifically, $L_4$ is zero if and only if $p_1 \geq \frac{\rho \Delta + c_1}{\eta_1}$, and is strictly decreasing in $p_1 < \frac{\rho \Delta + c_1}{\eta_1}$.} Similarly, if $c_1$ is not sufficiently small relative to $c_2$, then even without the concern for inducing the truthful report, the principal has to pay sufficiently large bonus to just to induce effort in $t = 1$. Therefore, agent with $y_1 = 2$ does not have much incentive to withhold the output in $t = 1$, i.e., $L_1$ is sufficiently close to zero.\footnote{More specifically, $L_1$ is zero if and only if $c_1 \geq \frac{\rho \Delta}{\eta_1-1}$, and is strictly decreasing in $\frac{c_1}{c_2} < \frac{\rho \Delta}{\eta_1-1}$.}

Given these observations on $L_1$, and Theorem 1, which tells us that $S(p_1, c_2) - C(p_1, c_2)$ is increasing in $p_1$ and $c_2$, we can state conditions under which the inequality (4) is likely to hold.

Suppose that $p_1$ is sufficiently large. We already know that $L_1$ is sufficiently small. Moreover, $L_2$ is sufficiently small because the ex-ante probability that the system gaming happens, $Pr(y_1 = 1)$, is pretty small. As a result, the losses arise from lack of commitment is decreasing in $p_1$, and hence (4) is likely to hold for a sufficiently large $p_1$.

When $c_2$ is sufficiently large, however, even though the termination contract is preferred over the linear contract, i.e., $S(p_1, c_2) > C(p_1, c_2)$, the principal may not prefer the quota contract over the linear contract for two reasons. The first reason is that $L_1$ can become large (especially when $c_1$ is small), making the cost of inducing truthful report to outweigh the net savings in agency cost. In contrast, if the relative size of $c_1$ is sufficiently high, then $L_1$ is sufficiently small, and hence the principal is likely to prefer the quota contract over the linear contract. The second reason is that the terms $L_2$ may be quite large relative to the size of $S(p_1, c_2) - C(p_1, c_2)$, especially if principal’s discounting factors $\delta$ is small. To be more precise, for a given $p_1$ and $c_2 > c_2^{EC}(p_1)$, there exists a $\delta(p_1, p_1, c_2) < 1$ such that $S(p_1, c_2) - C(p_1, c_2) > L_2(p_1, c_1, c_2)$ if and only if $\delta > \delta(p_1, c_2)$.\footnote{Note that

\[
\frac{\partial (S(p_1, c_2) - C(p_1, c_2) - 2p_1(1 - p_1)(1 - \delta))}{\partial \delta} = (1 - p_1) \frac{2\Delta p_1 - (1 - p_1) \Delta^2 + (1 - p_1)p_1 c_2}{\Delta} > (1 - p_1) \frac{\Delta (2p_1 - (1 - p_1) \Delta)}{\Delta} > 0.
\]

Moreover, $\lim_{\delta \to 1} (S(p_1, c_2) - C(p_1, c_2) - 2p_1(1 - p_1)(1 - \delta)) > 0$ by assumption. Therefore, we have the required result.}

We thus arrive at the following conclusion.

**Theorem 4** Suppose that $\rho \approx \delta$.

1. For any $c_1$ and $c_2$, there exists a $p_1(c_1, c_2) > p_1^{EC}(c_2)$ such that the principal chooses the quota contract over the linear contract if and only if $p_1 \geq p_1(c_1, c_2)$.\footnote{Note that

\[
\frac{\partial (S(p_1, c_2) - C(p_1, c_2) - 2p_1(1 - p_1)(1 - \delta))}{\partial \delta} = (1 - p_1) \frac{2\Delta p_1 - (1 - p_1) \Delta^2 + (1 - p_1)p_1 c_2}{\Delta} > (1 - p_1) \frac{\Delta (2p_1 - (1 - p_1) \Delta)}{\Delta} > 0.
\]

Moreover, $\lim_{\delta \to 1} (S(p_1, c_2) - C(p_1, c_2) - 2p_1(1 - p_1)(1 - \delta)) > 0$ by assumption. Therefore, we have the required result.}
2. For any \( p_1, c_2, \) and \( \delta > \bar{\delta}(p_1, c_2) \), there exists a \( \bar{c}_1(p_1, c_2) \) and the principal chooses the quota contract over the linear contract if and only if \( c_1 \geq \bar{c}_1(p_1, c_2) \) and \( c_2 \geq c_2^{RP}(p_1, c_1) \).

**Proof.** In the Appendix.

3 **Optimality of Quota Contract**

In this section, we provide sufficient conditions that ensure the optimal quota contract is an optimal renegotiation-proof contract. We have seen above that a benefit of using the quota contract is that by inducing system gaming for the agent with \( y_1 = 1 \), it helps the principal to commit to credibly punish the agent.

However, other forms of contract may also be able to achieve the same goal. For example, the principal may find it beneficial to induce the system gaming by the agent with \( y_1 = 2 \) alone, or both \( y_1 = 1 \) and 2. We show below that the optimal renegotiation-proof contract is the best way to minimize the cost associated with system gaming, while still keeping the cost of renegotiation high enough.

To see this, recall that the optimal renegotiation-proof contract induces the same effort profile as the termination contract, an optimal contract under the full commitment. If the principal wants to induce this effort profile without using quota contract, then she has to use a contract that either induces the agent with \( y_1 = 2 \) to game the system along with \( y_1 = 0 \); or induces the agent to game the system irrespective of his performance. In either way, it raises the cost of system gaming even if such a contract is renegotiation-proof. That is, the quota-contract is the least costly way to induce the effort profile.

For the same reason, if the principal uses a contract that induces an effort profile that is different from one under the termination contract, she cannot lower the cost of system gaming. Moreover, the optimality of the termination contract under the full commitment implies that inducing such an effort profile is inefficient from the principal’s perspective.

These observations lead us to conclude that when the optimal renegotiation-proof quota contract outperforms the linear contract, it also likely to outperform any other renegotiation-proof contract.

**Theorem 5** Suppose that \( \rho \approx \delta \).

1. For any \( c_1 \) and \( c_2 \), there exists a \( p_1^* (c_1, c_2) \geq \bar{p}_1 (c_1, c_2) \) and the optimal renegotiation-proof quota contract is an optimal contract if \( p_1 \geq p_1^* (c_1, c_2) \).

2. For any \( p_1, c_2 \), and \( \delta > \bar{\delta}(p_1, \rho, c_2) \) there exists a \( c_1^* (p_1, c_2) \geq \bar{c}_1 (p_1, c_2) \) and the optimal renegotiation-proof quota contract is an optimal contract if \( c_1 \geq c_1^* (p_1, c_2) \) and \( c_2 \geq c_2^{RP} (p_1, c_1) \).

**Proof.** In the Appendix.
4 Concluding Remarks

By identifying a novel benefit of incentive-system gaming, we shed light on the widespread use of the quota contract.

Our model allows the principal to offer any general renegotiation mechanism, which may not be the most realistic assumption. An alternative specification is to allow the principal ONLY to improve existing bonus payments (but cannot offer a menu of contracts from which the agent can pick one). This specification makes the quota contract renegotiation-proof for a wider range of parameters, and thus strengthen the result that quota contract can outperform linear contract.15)

There are a few natural extensions that could be profitably studied. First, we consider an employment relationship that lasts for only two periods. It is interesting and important to extend the analysis to an employment relationship that lasts for more than two periods. A quota contract in such an environment allows the principal to temporally and probabilistically suspend the agent’s production and thus lower the overall agency rent. We thus believe that our main result will still hold: a quota contract can outperform linear contracts, particularly when moral hazard problem is severe.

Another natural extension would be to study the possibility of “pulling in.” Should the agent can both pull in and push out, the agent who marginally fails to meet the quota may choose to pull in rather than push out. Conditional on that the agent failing to meet the quota, however, it is quite unlikely that the agent has attempted pulling in. Therefore, the principal can maintain a high quota in the next period for those who failed to meet the quota. This in turn induces the pushing out, and hence enables the principal to punish the agent whose performance is poor. As a result, the quota contract can outperform the linear contract. Formal and careful analysis of such environment awaits further study.

5 Appendix

Proof of Theorem 1

We only need to consider four contracts: (i) the contract that always induces effort in \( t = 1 \) and \( t = 2 \) for all \( y_1 \); (ii) the contract that always induces effort in \( t = 1 \) and effort in \( t = 2 \) if and only if \( y_1 = 1 \) or 2; (iii) the contract that induces effort in \( t = 2 \) if and only if \( y_1 = 2 \); and (iv) the contract that induces effort in the second round of \( t = 1 \) if and only if the first round is success, and effort in \( t = 2 \) if and only if \( y_1 = 1 \) or 2.

15) In fact, the quota contract is never renegotiation-proof if \( c_1 = c_2 \). Nevertheless, we obtain a similar result if the principal is only allowed to improve the existing contract, instead of being allowed to make any new proposal. More specifically, if the original contract specified \( b_2(r_2; r_1) \) as the bonus for the second period, then the offer the principal can make a new offer \( \tilde{b}_2(r_2; r_1) \) if and only if \( \tilde{b}_2(r_2; r_1) \geq b_2(r_2; r_1) \).
We start with the optimal contract with property-(i). By assumptions on $c_1$ and $c_2$, it is straightforward to see that if the principal induces effort in $t = 2$ irrespective of $y_{1t}$, then linear contract, with $\alpha_1 = c_1/\Delta$ and $\alpha_2 = c_2/\Delta$, is the optimal contract. The principal’s payoff $\Pi^{LC}$ by using the linear contract is

$$\Pi^{LC} = p_1^2 \left( 2 \times \left( 1 - \frac{c_1}{\Delta} \right) + \delta p_1 \left( 1 - \frac{c_2}{\Delta} \right) \right) + 2p_1 (1 - p_1) \left( \left( 1 - \frac{c_1}{\Delta} \right) + \delta p_1 \left( 1 - \frac{c_2}{\Delta} \right) \right) + (1 - p_1)^2 \delta p_1 \left( 1 - \frac{c_2}{\Delta} \right) = 2p_1 \left( 1 - \frac{c_1}{\Delta} \right) + \delta p_1 \left( 1 - \frac{c_2}{\Delta} \right).$$

Next, we consider the contract with the property-(ii). It is straightforward to see that $b_2(1; 2) = b_2(1; 1) = c_2/\Delta$, and $b_2(1; 0) = b_1(0; \cdot) = 0$. Under such a contract, the agent exerts effort in the second round in $t = 1$ following the failure in the first round if and only if

$$p_1 b_1(1) + \rho \left( \frac{p_0 c_2}{p_1 - p_0} \right) - c_1 \geq p_0 \left( b_1(1) + \rho \left( \frac{p_0 c_2}{p_1 - p_0} \right) \right) \quad \Rightarrow \quad b_1(1) \geq \frac{c_1 - \rho (p_1 - \Delta) c_2}{\Delta}. \quad (5)$$

The agent exerts effort in the second round in $t = 1$ following the success if and only if

$$p_1 b_1(2) + \rho \left( \frac{p_0 c_2}{p_1 - p_0} \right) - c_1 \geq p_0 b_1(2) + \rho \left( \frac{p_0 c_2}{p_1 - p_0} \right),$$

which can be simplified to

$$b_1(2) \geq \frac{c_1}{p_1 - p_0} + b_1(1) = \frac{2c_1 - \rho (p_1 - \Delta) c_2}{\Delta}. \quad (6)$$

The agent exerts effort in the first round if and only if

$$\Delta \left( p_1 b_1(2) + (1 - p_1) b_1(1) + \rho \left( \frac{(p_1 - \Delta) c_2}{\Delta} \right) - c_1 \right) - \Delta \left( p_1 b_1(1) + \rho \left( \frac{(p_1 - \Delta) c_2}{\Delta} \right) - c_1 \right) \geq c_1$$

which is simplified to

$$p_1 b_1(2) + (1 - 2p_1) b_1(1) \geq \frac{c_1 - (1 - p_1) \rho (p_1 - \Delta) c_2}{\Delta}. \quad (7)$$

Note that (5) and (6) together imply (7). Therefore, the principal’s payoff under
this termination contract is

$$\Pi^T = p_1^2 \left( 2 - \frac{2c_1 - \rho(p_1 - \Delta)c_2}{\Delta} + \delta p_1 \left( 1 - \frac{c_2}{\Delta} \right) \right)$$

$$+ 2p_1 (1 - p_1) \left( 1 - \frac{c_1 - \rho(p_1 - \Delta)c_2}{\Delta} \right) + \delta p_1 \left( 1 - \frac{c_2}{\Delta} \right) + (1 - p_1)^2 \delta (p_1 - \Delta).$$

For the contract with property-(iii), we know that it is optimal to set $b_2 (1; r_1) = 0$ for $r_1 = 0, 1$. Then, the agent never exerts effort in the second round of $t = 1$ if he fails in the first round. The agent exerts effort in the second round in $t = 1$ following the success if and only if

$$\Delta \left( b_1 (2) + \rho \left( \frac{(p_1 - \Delta)c_2}{\Delta} \right) \right) \geq c_1$$

$$b_1 (2) \geq \frac{c_1 - \rho(p_1 - \Delta)c_2}{\Delta}.$$

The agent exerts effort in the first round if and only if

$$\Delta \left( p_1 \left( b_1 (2) + \rho \left( \frac{(p_1 - \Delta)c_2}{\Delta} \right) \right) - c_1 \right) \geq c_1$$

$$b_1 (2) \geq \max \left\{ \frac{(1 + \Delta) c_1 - \rho p_1 (p_1 - \Delta)c_2}{p_1 \Delta}, 0 \right\}.$$

Since $\frac{(1 + \Delta) c_1 - \rho p_1 (p_1 - \Delta)c_2}{p_1 \Delta} > \frac{c_1 - \rho(p_1 - \Delta)c_2}{\Delta}$, it is optimal to set $b_1 (2) = \frac{(1 + \Delta) c_1 - \rho p_1 (p_1 - \Delta)c_2}{p_1 \Delta}$. The principal’s payoff by offering this contract is given by

$$\Pi^{(iii)} = p_1^2 \left( 2 - \frac{(1 + \Delta) c_1 - \rho p_1 (p_1 - \Delta)c_2}{p_1 \Delta} \right. \left. + \delta p_1 \left( 1 - \frac{c_1}{\Delta} \right) \right)$$

$$+ p_1 (1 - p_1) (1 + \delta (p_1 - \Delta)) + (1 - p_1) (p_1 - \Delta) (1 + \delta).$$

First, we show that $\Pi^T > \Pi^{(iii)}$. To see this, define $\Delta \Pi^{(iii)} (c_1, c_2) \equiv \Pi^T - \Pi^{(iii)}$. Then

$$\frac{\partial \Delta \Pi^{(iii)}}{\partial c_1} = -p_1 \frac{1 - \Delta}{\Delta} < 0,$$

$$\frac{\partial \Delta \Pi^{(iii)}}{\partial c_2} = -2p_1 \frac{(\delta - \rho) p_1 + \rho \Delta (1 - p_1)}{\Delta} < 0.$$

Since $\Delta \Pi^{(iii)} (c_1, c_2) > 0$, we have $\Pi^T > \Pi^{(iii)}$.

Now we see the contract with property (iv). The principal sets $b_2 (1; 2) = b_2 (1; 1) = c_2 / \Delta$, $b_2 (1; 0) = 0$, and $b_1 (1) = 0$.

The agent exert effort in the first round in $t = 1$ if only if

$$\Delta \left( p_1 b_1 (2) - c_1 + \rho \frac{(p_1 - \Delta)c_2}{\Delta} \right) - (p_1 - \Delta) \rho \frac{(p_1 - \Delta)c_2}{\Delta} \geq c_1$$

$$b_1 (2) \geq \frac{(1 + \Delta) c_1 - \rho(p_1 - \Delta) (1 - (p_1 - \Delta)) c_2}{\Delta p_1}.$$
The agent exerts effort in the second round after success because \(b_1(2) \geq c_1/\Delta\), but does not because

\[
p_1 \rho \frac{(p_1 - \Delta) c_2}{\Delta} - c_1 \leq (p_1 - \Delta) \rho \frac{(p_1 - \Delta) c_2}{\Delta}
\]

follows from \(c_1 > \rho (p_1 - \Delta) c_2\). Therefore, the principal’s payoff under full commitment for inducing the effort profile is

\[
\Pi^{(iv)} = p_1^2 \left( 2 - \frac{(1 + \Delta) c_1 - \rho (p_1 - \Delta) (1 - (p_1 - \Delta)) c_2}{\Delta p_1} \right) + (1 - p_1) (p_1 + (p_1 - \Delta)) + \delta \left( (1 - (1 - p_1)(1 - (p_1 - \Delta))) p_1 \left( 1 - \frac{c_2}{\Delta} \right) + (1 - p_1) (1 - (p_1 - \Delta)) (p_1 - \Delta) \right)
\]

Define \(\Delta \Pi^{(iv)}(c_1, c_2) \equiv \Pi^T - \Pi^{(iv)}\). Then

\[
\frac{\partial \Delta \Pi^{(iv)}(c_1, c_2)}{\partial c_1} = -\frac{p_1}{\Delta} (1 - \Delta) < 0,
\]

Also note that \(\Delta \Pi^{(iv)}(c_1, c_2)\) is linear in \(c_2\), and

\[
\Delta \Pi^{(iv)}(\tilde{c}_1, \tilde{c}_2) = \rho (p_1 - \Delta) \Delta (1 - \Delta) > 0, \text{ and }
\Delta \Pi^{(iv)}(\tilde{c}_1, 0) = \delta (1 - p) \Delta^2 > 0.
\]

Therefore, we are done if we show \(\Delta \Pi^{(iv)}(c_1, c_2) > \Delta \Pi^{(iv)}(\tilde{c}_1, c_2) > 0\).

Now we now show that \(\Pi^{LC} > \Pi^T\) if and only if \(p_1 > p_1^{FC}\) for some \(p_1^{FC} \in (0, 1)\).

Note that for \(\Delta \Pi \equiv \Pi^{LC} - \Pi^T\),

\[
\frac{\partial \Delta \Pi}{\partial p_1} = 2p_1 \left( \frac{\rho (p_1 - \Delta) c_2}{\Delta} \right) + p_1^2 \frac{\rho c_2}{\Delta} + 2(1 - 2p_1) \left( \frac{\rho (p_1 - \Delta) c_2}{\Delta} \right)
+ 2p_1 (1 - p_1) \left( \frac{\rho c_2}{\Delta} \right) + 2(1 - p_1) \delta \left( \Delta - \frac{p_1 c_2}{\Delta} \right) + (1 - p_1)^2 \delta \frac{c_2}{\Delta}
\geq 0,
\]

and \(\lim_{p_1 \to 1} \Delta \Pi > 0\). Therefore, we have the required result.

Lastly, we show that \(\Pi^{LC} > \Pi^T\) if and only if \(c_2 < c_2^{FC} \equiv \frac{\delta (1 - p)^2 \Delta^2}{p_1 (\rho (p_1 - \Delta)(2 - p_1) + \delta (1 - p))}\).

To see this, let \(\Delta \Pi = \Pi^T - \Pi^{LC}\). Then

\[
\frac{\partial \Delta \Pi}{\partial c_2} = p_1^2 \left( \frac{\rho (p_1 - \Delta)}{\Delta} \right) + 2p_1 (1 - p_1) \left( \frac{\rho (p_1 - \Delta)}{\Delta} \right) + (1 - p_1)^2 \delta \frac{p_1}{\Delta} > 0
\]

and \(\Delta \Pi(c_2) = 0\). We thus are done. Q.E.D.
Proof of Lemma 1

We derive conditions under which the quota contract is renegotiation-proof. Let us consider the principal’s problem at the beginning of \( t = 2 \) following a report \( r_1 = 0 \). If the principal does not renegotiate, his expected payoff is

\[
\Pi^Q_2 \equiv \frac{2p_1 (1 - p_1)}{2p_1 (1 - p_1) + (1 - p_1)^2} (1 + p_1 (1 - \beta)) + \frac{(1 - p_1)^2}{2p_1 (1 - p_1) + (1 - p_1)^2} p_0
\]

In our simple setup, the only benefit to the principal from renegotiation is to elicit effort from worker with \( y_1 = 0 \). To this end, it is without loss to focus on menu of contract \( \{b^y_2 (r_2)\}_{y_1=0,1} \) of the following form:

\[
\begin{align*}
&b^0_2 (0; 0) = 0, b^0_2 (1; 0) = b^0_2 (2; 0) = c_2 / \Delta; \\
&b^1_2 (0; 0) = b^1_2 (1; 0) = 0, b^1_2 (2; 0) = \tilde{b}.
\end{align*}
\]

The value of \( \tilde{b} \) has to be chosen such that (i) agent with \( y_1 \) chooses \( b^y_2 (\cdot) \); (ii) agent with \( y_1 = 1 \) exerts effort in \( t = 2 \); and (iii) \( \tilde{b} \geq \beta \). These requirements translate into

\[
\tilde{b} \geq \max \left\{ \frac{(1 + \Delta) c_2}{p_1 \Delta}, \beta \right\} = \tilde{b} (\beta).
\]

Therefore the principal’s expected payoff from renegotiation is

\[
\Pi^R_2 \equiv \frac{2p_1 (1 - p_1)}{2p_1 (1 - p_1) + (1 - p_1)^2} (1 + p_1 (1 - \tilde{b} (\beta))) + \frac{(1 - p_1)^2}{2p_1 (1 - p_1) + (1 - p_1)^2} p_1 \left( 1 - \frac{c_2}{\Delta} \right).
\]

On comparison of \( \Pi^Q_2 \) and \( \Pi^R_2 \), we can conclude that for a given \( \beta \), the principal find the renegotiation unprofitable if and only if

\[
\frac{2p_1^2 (1 - p_1)}{2p_1 (1 - p_1) + (1 - p_1)^2} (\tilde{b} (\beta) - \beta) \geq \frac{(1 - p_1)^2}{2p_1 (1 - p_1) + (1 - p_1)^2} (\Delta - p_1 c_2 / \Delta).
\]

Costs of Renegotiation

Gains from Renegotiation

Note that there exists a \( \tilde{\beta} \), and (8) is satisfied if and only if \( \beta \leq \tilde{\beta} \).

Next, let’s work out the incentive constraints that ensure the agent chooses the stated effort profile. To induce effort in the second round of \( t = 1 \) following the failure in the first round, the following inequality has to hold

\[
\rho p_1 (p_1 \beta - c_2) - c_1 \geq \rho (p_1 - \Delta) (p_1 \beta - c_2).
\]

Therefore, the agent exert effort in the second round of \( t = 1 \) following the failure if and only if

\[
\beta \geq \beta^* = \frac{p \Delta c_2 + c_1}{p_1 \rho \Delta}.
\]

(9)
To induce the effort in the second round of \( t = 1 \) following the success in the first round, \( b_1(2) \equiv B \) and \( b_2(2; y_1) \equiv \beta \) have to satisfy the following inequality:

\[
p_1 \left( B + \frac{\rho \left( p_1 - \Delta \right) c_2}{\Delta} \right) + (1 - p_1) \rho (p_1 \beta - c_2) - c_1 \\
\geq p_0 \left( B + \frac{\rho \left( p_1 - \Delta \right) c_2}{\Delta} \right) + (1 - p_0) \rho (p_1 \beta - c_2).
\]

Thus, for a given \( \beta \), the agent exert effort in the second round if and only if

\[
B \geq \frac{\rho (p_1 - \Delta) c_2}{\Delta} + \rho (p_1 \beta - c_2) \equiv B^E (\beta).
\]

Now we show that if \( \beta \geq \beta^* \), and \( B \geq B^E (\beta^*) \), then the agent’s incentive compatibility constraint for effort in the first round of \( t = 1 \) is satisfied. To see this, suppose that \( \beta = \beta^* \). Then the continuation payoff following failure in the first stage of period 1 is \( p_1 \rho (p_1 \beta^* - c_2) - c_1 = p_0 c_1 / \Delta \). Also at \( \beta = \beta^* \) and \( B = B^E (\beta^*) \), the continuation payoff following success in the first stage of period 1 is \( p_1 \left( B^E + \frac{\rho (p_1 - \Delta) c_2}{\Delta} \right) + (1 - p_1) \rho (p_1 \beta^* - c_2) - c_1 = \frac{1 + p_1 - \Delta}{\Delta} c_1 \). The difference is \( c_1 / \Delta \), and the agent is indifferent between exerting effort or not in the first round of \( t = 1 \). Since \( \beta \geq \beta^* \) and \( B^E (\beta) \) is increasing in \( \beta \), we know that incentive constraint for effort at the first round of \( t = 1 \) does not bind.

The final incentive constraint concerns inducing agent with \( y_1 = 2 \) to report truthfully. The payoff of reporting \( r_1 = 2 \) is weakly higher than reporting \( r_1 = 0 \) if and only if \( B + \rho \frac{(p_1 - \Delta) c_2}{\Delta} \geq \rho \beta \). Thus, for a given \( \beta \), the agent with \( y_1 = 2 \) reports truthfully if and only if

\[
B \geq \rho \left( \beta - \rho \frac{(p_1 - \Delta) c_2}{\Delta} \right) \equiv B^T (\beta).
\]

Therefore, the principal would find it optimal to set \( \beta \) as low as possible, since doing so would not tighten the incentive constraints and renegotiation-proofness constraint (note that both \( B^E (\beta) \) and \( B^T (\beta) \) are decreasing in \( \beta \). Thus, we can focus on quota contract with \( \beta = \beta^* \) and \( B = B^* (\beta^*) \equiv \max \{ B^E (\beta), B^T (\beta) \} \).

That is,

\[
B^* (\beta^*) = \begin{cases} 
B^E (\beta^*) = \frac{2c_1 - \rho (p_1 - \Delta) c_2}{\Delta} & \text{if } p_1 \geq \hat{p}_1 \equiv \frac{p_1 c_1 + \Delta}{2c_1} \\
B^T (\beta^*) = \frac{\rho (\Delta - (p_1 - \Delta) p_1) c_2 + c_1}{\rho (p_1 - \Delta)} & \text{if } p_1 < \hat{p}_1
\end{cases}
\]

Therefore, we can conclude that a quota contract is renegotiation-proof if and only if

\[
\frac{2p_1^2 (1 - p_1)}{2p_1 (1 - p_1) + (1 - p_1)^2} \left( \frac{c_2 \rho - c_1}{p_1 p_0 \Delta} \right) \geq \frac{(1 - p_1)^2}{2p_1 (1 - p_1) + (1 - p_1)^2} (\Delta - p_1 c_2 / \Delta). \tag{10}
\]

Q.E.D.
Proof of Theorem 3

Define \( RC(p_1, c_1, c_2) \equiv 2p_1 (c_2 - c_1 / \rho) - (1 - p_1)(\Delta^2 - p_1c_2) \). Then, (10) holds if and only if \( RC(p_1, c_1, c_2) \geq 0 \). Note that \( RC(p_1, c_1, c_2) \) is strictly increasing in \( p_1 \). This is because \( c_2 > c_1 / \rho \) and \( c_2 < \Delta^2 / p_1 \) by assumption, and hence

\[
\frac{\partial RC(p_1, c_1, c_2)}{\partial p_1} = 2 \left( c_2 - \frac{c_1}{\rho} \right) + (\Delta^2 - p_1c_2) + (1 - p_1)c_2 > 0.
\]

Moreover, \( \lim_{p_1 \to 1} RC(p_1, c_1, c_2) = \left( c_2 - \frac{c_1}{\rho} \right) > 0 \). Therefore, \( RC(p_1, c_1, c_2) > 0 \) if and only if \( p_1 \geq \frac{\rho c_1}{\Delta^2 - \frac{c_1}{\rho}} \frac{\rho c_1 + \Delta^2 (1 - p_1)}{\rho c_1 + \Delta^2 (1 - p_1)} \).

Next, note that \( RC(p_1, c_1, c_2) \) is increasing in \( c_2 \). Therefore, \( RC(\cdot) > 0 \) if and only if \( c_2 > \frac{\rho c_1 + \Delta^2 (1 - p_1)}{\rho c_1 + \Delta^2 (1 - p_1)} \). Q.E.D.

Proof of Theorem 4

To see (i), note that if \( p_1 > \hat{p}_1 = \frac{\rho c_1}{\Delta^2 - \frac{c_1}{\rho}} \), the \( L(p_1, c_1, c_2) = 2p_1 (1 - p_1)(1 - \delta) + 2p_1 (1 - p_1) \left( (\delta - \rho) \left( \frac{c_1 - c_2}{\rho \Delta} \right) \right) \), which is decreasing in \( p_1 \) when \( \rho = \delta \). Moreover, \( \lim_{p_1 \to 1} L(p_1, c_1, c_2) = 0 \). Hence, we have the required result. To see (ii), note that \( \Lambda \equiv S(p_1, c_2) - C(p_1, c_2) - 2p_1 (1 - p_1)(1 - \delta) > 0 \) by assumption. Moreover, \( \Lambda \) does not depend on \( c_1 \). On the other hand, \( p_1 \left( \max \left\{ \frac{\rho c_1 + \Delta^2 (1 - p_1)}{\rho c_1 + \Delta^2 (1 - p_1)} \right\} \right) \) is strictly decreasing in \( c_1 \) when \( c_1 < \frac{\rho c_1}{\Delta^2 - \frac{c_1}{\rho}} \). Therefore, there exists a \( \bar{c}_1(p_1, c_2) \) and \( \Lambda > p_1 \left( \max \left\{ \frac{\rho c_1 + \Delta^2 (1 - p_1)}{\rho c_1 + \Delta^2 (1 - p_1)} \right\} \right) \) if and only if \( c_1 > \bar{c}_1(p_1, c_2) \). Q.E.D.

Proof of Theorem 5

For the simplicity of exposition, we refer the optimal renegotiation-proof contract as the quota contract throughout the proof. Also we use \( \Pi^Q \) to represent the principal’s payoff under the quota contract.

First, we prove that there is no renegotiation-proof contract that induces the same effort profile as the termination contract under full commitment, and gives a higher payoff than the quota contract. That is, we prove that the quota contract is a least costly way to implement the effort profile of the termination contract under the full commitment. For this end, we only need to consider the following two contract forms

contract-(i):
\[
\begin{align*}
b_1(0) &= 0, b_1(1) = B; b_1(2) = B > 0 \\
b_2(3;0) &\equiv \beta; b_2(y_2;0) = 0 \text{ for } y_2 = 0, 1, 2 \\
b_2(1;1) &= c_2 / \Delta; b_2(0;1) = 0.
\end{align*}
\]

contract-(ii):
\[
\begin{align*}
b_1(0) &= b_1(1) = b_1(2) = 0 \\
b_2(3;0) &\equiv B; b_2(2;0) = \beta; b_2(y_2;0) = 0 \text{ for } y_2 = 0, 1
\end{align*}
\]

The explicit expression is \( \bar{c}_1(p_1, c_2) = \frac{\Delta^2 - \frac{c_1}{\rho}}{\rho c_1 + \Delta^2 (1 - p_1)} \).

\[\frac{\rho c_1}{\Delta^2 - \frac{c_1}{\rho}}.\]
Under this contract, the agent always reports their performance, i.e., the agent with yields payoff $\Pi - \max\{0, \beta - c_2\}$. To see that contract-(i) is not renegotiation-proof, note that effort in the second round of $t = 1$ following failure in the first round requires $B \geq \frac{c_1 - \rho (p_1 - \Delta) c_2}{\rho p_1 \Delta}$. To induce effort in the second round of $t = 1$ following success in the first round requires $\Delta \left( p_1 \beta - \left( \frac{p_1}{\Delta} \right) c_2 \right) - B \geq c_1$. This in turn implies $\beta \geq \frac{2c_1 + \rho c_2 \Delta}{p_1 \rho \Delta}$. For the agent with $y_1 = 2$ to NOT report $r_1 = 2$ in $t = 1$, we have to have $\rho (p_1 \beta - c_2) \geq B + \rho \frac{c_2}{\Delta}$, which implies $\beta \geq \frac{2c_1 \rho c_2 \Delta + c_1 - \rho p_1 p_2 \rho c_2}{\rho p_1 \Delta}$.

To induce the effort in the first round of $t = 1$, we need

$$\Delta \left( p_1 \rho \left( p_1 \beta - c_2 \right) + (1 - 2p_1) \left( \frac{p_1}{\Delta} - \Delta \right) c_2 \right) \geq c_1,$$

which implies $\beta \geq \frac{2c_1 + \rho c_2 \Delta}{p_1 \rho \Delta}$. Since $\frac{2c_1 + \rho c_2 \Delta}{p_1 \rho \Delta} = \max \left\{ \frac{2c_1 + \rho c_2 \Delta}{p_1 \rho \Delta}, \frac{2p_1 \rho c_2 \Delta + c_1 - \rho p_1 p_2 \rho c_2}{p_1 \rho \Delta} \right\}$, the principal sets $\beta = \frac{2c_1 + \rho c_2 \Delta}{p_1 \rho \Delta}$ and $B = \frac{c_1 - \rho (p_1 - \Delta) c_2}{\rho p_1 \Delta}$. However, under such a contract, the continuation payoff of the agent with $y_1 = 2$ is

$$\frac{2c_1 + \rho c_2 \Delta}{\rho \Delta} - c_2 = \frac{2c_1}{\rho \Delta} > c_2.$$

Therefore, at $t = 2$, the principal can screen the agent $y_1 = 0$ from $y_1 = 2$ without any additional cost, i.e., this contract is not renegotiation-proof.

Next, we consider contract-(ii). The effort in $t = 2$ for the agent with $y_1 = 2$ requires $B \geq \frac{p_1}{\Delta} + \beta$. The effort in the second round of $t = 1$ following failure in $t = 1$ requires $\beta \geq \frac{c_1 + \Delta c_2}{\rho p_1 \Delta}$. Therefore, the principal sets

$$\beta \geq \frac{c_1 + \Delta c_2}{\rho p_1 \Delta} \quad \text{and} \quad B \geq \frac{c_1 + (\Delta + \rho p_1) c_2}{\rho p_1 \Delta}.$$

It is straightforward to check that $\beta B \geq B^* (\beta^*)$, where $B^* (\beta^*)$ is the payment under the quota contract specified in (2). Hence, when the quota-contract is renegotiation-proof, the principal always pays “more,” i.e., she does not use this contract to induce the same effort profile as the termination contract.

Next, we consider the effort profile in which the principal induces effort (i) in the first round in $t = 1$, (ii) in the second round in $t = 1$ if and only if the first round is success, and (iii) in $t = 1$ if and only if $y_1 = 1$ or 2. To induce this effort profile, we only need to consider the following contract, which the principal yields payoff of $\Pi^1$.

$$b_1 (0) = b_1 (1) = b_1 (2) = 0$$
$$b_2 (3; 0) \equiv \beta; b_2 (2; 0) = B, b_2 (y_2; 0) = 0 \text{ for } y_2 = 0, 1$$

Under this contract, the agent always reports $r_1 = 0$, irrespective of his true performance, i.e., the agent with $y_1 \neq 0$ is induced to game the system. This
implies the cost of system gaming under this contract (in comparison to the linear contract) is always larger than the cost of system gaming under the quota contract where only the agent with $y_1 = 1$ games the system. Therefore, when the conditions of the theorem are satisfied,

$$\Pi^Q - \Pi^{LC} \approx \Pi^{TC} - \Pi^{LC} - 2p_1 (1 - p_1) (1 - \delta)$$

Moreover, note that from proof of Theorem 1, we know under the full commitment, the termination contract outperforms the optimal contract that induces this effort profile. Therefore,

$$\Pi^Q - \Pi^{LC} > \Pi^{TC} - \Pi^{LC} - 2p_1 (p_1 (2p_1 + (1 - p_1)) + (1 - p_1) (p_1 - \Delta_1))$$

Lastly, consider the effort profile in which the principal induces effort (i) in the first round in $t = 1$, (ii) in the second round in $t = 1$ if and only if the first round is success, and (iii) in $t = 1$ if and only if $y_1 = 2$. For a contract that induces this effort profile to be renegotiation-proof, the agent always has to report $r_1 = 0$, irrespective of his true performance, i.e., the agent with $y_1 \neq 0$ is induced to game the system. Therefore, we only need to consider the following contract, which the principal yields the payoff of $\Pi^{II}$.

$$b_1 (0) = b_1 (1) = b_1 (2) = 0$$
$$b_2 (3; 0) \equiv \beta; b_2 (y_2; 0) = 0 \text{ for } y_2 = 0, 1, 2$$

Note that the cost of system gaming under this contract (in comparison to the linear contract) is always larger than the cost of system gaming under the quota contract where only the agent with $y_1 = 1$ games the system. Moreover, from proof of Theorem 1, we know that under the full commitment, the termination contract outperforms the optimal contract that induces this effort profile. Therefore, the same argument as above leads us to conclude that $\Pi^Q - \Pi^{LC} > \Pi^{II} - \Pi^{LC}$ when the conditions of the theorem are met. Q.E.D.

References


