Optimality of Quota Contract

Pak-Hung Au*, Keiichi Kawai‡

May 8, 2015

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, the agent can manipulate the timing of closing a deal and/or reporting earnings. Because of this 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 for the principal if she has a contract commitment problem. Moreover, the benefit of a quota contract arises precisely from inducing the agent to game the incentive system.

1 Introduction

A quota-based contract – a form of incentive contracts 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 contracts, especially for salespeople and executives.[1)] 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., manipulation of the timing of business by the agent to increase his overall compensation.

Empirical and anecdotal evidence abounds for system gaming activities, and the associated costs they impose on the firms. For example, by looking into

*E-mail: PHAU@ntu.edu.sg. Division of Economics, Nanyang Technological University

†E-mail: k.kawai@unsw.edu.au. School of Economics, The University of New South Wales

‡The authors thank Jimmy Chan, Stephen Chiu, Dino Gerardi, Toomas Hinnosaar, Richard Holden, Anton Kolotilin, Hongyi Li, Bart Lipman, Satoru Takahashi, Bingyong Zheng, and seminar participants at Collegio Carlo Alberto, Nanyang Technological University, National University of Singapore, University of New South Wales and University of Queensland for useful discussions and valuable comments. The first author greatly acknowledges the financial support from Shanghai University of Finance and Economics, and Nanyang Technological University. The second author greatly acknowledges the financial support from Collegio Carlo Alberto, Nanyang Technological University and University of Queensland.

1)According to **?**, quota-based compensation is one of the most consistent features in the sales industry. Specifically, **?** notes that over 80% of salespeople are compensated according to a quota scheme.

executives' accrual decision, **?** finds that executives do not report 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, **?** 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 current-period quota; or "push out" business to the next fiscal period if they fall too far behind the current-period quota. The timing manipulation by agents generates a spike in reported output at the end of each period, and a dip at the beginning of each period, resulting in undesirable output fluctuation for the firm. In a similar vein, **?** 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 costs by deliberately inducing agents to game the system. Our theory thus sheds light on the widespread use of quota-based contracts in various industries.

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. The agent is protected from limited liability: he must be paid a nonnegative wage in every period. In the first period, 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 the first period and report them in the second period, when 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 the second period, which imposes a direct time-discounting cost the principal. At first glance, the principal is always hurt by agents' system gaming, so a quota-based contract is dominated by a linear contract.

To understand our finding, note that in a dynamic moral hazard environment with limited liability, the principal can motivate the agent via both current incentives (immediate bonus payments for good performance) and fu-

[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 saleperson 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.

ture incentives (threat of lower future incentive payments or even inefficient termination of the relationship following poor performance). Although using future incentives is costly to the principal as valuable agent effort is forgone, the threat of punishment for poor performance help motivate the agent. As a result, future incentives can save agency payments, and the optimal long-term contract typically combines the use of both types 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 a bad performance, the principal has every incentive to replace the existing contract that punishes the agent's poor **past** performance (which results in low subsequent effort), with one that rewards the agent for good **future** performance (which results in high 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 modeled as the principal making a take-it-or-leave-it offer to the agent.[4)]

Below, we provide the intuition of why a quota-based contract helps making the punishment credible. Consider the following long-term contract: (i) a first-period bonus is paid if and only if the specified 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 the 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, a 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 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 her contract-commitment ability. This is possible because it induces the agent to game the incentive system, thus creating an endogenous 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 can enjoy the savings in overall agency

[4)] 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, **?** and **?**.

costs in comparison to contracts that do not induce system gaming, such as 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 the agent's system gaming activities as a dysfunctional response to quota-based contracts (see **?**), our theory proposes that such gaming activities can be beneficial to the principal, and she may deliberately design a contract and/or work environment in order to encourage such activities.

In our analysis, we provide sufficient conditions under which the optimal renegotiation-proof contract takes a quota form. Loosely speaking, these conditions ensure that (i) the optimal quota contract is renegotiation-proof; (ii) the optimal quota contract implements the effort profile in the optimal full-commitment contract; and (iii) the direct cost to the principal resulting from agent's system gaming is small.

Several papers have proposed explanations for the use of quota contracts. **?** shows the optimality of quota contract in a static moral hazard setting. **?** shows that a quota contract is optimal if the agent is expectation-based loss averse. **?** suggest that convex payment scheme can attract over-confident workers, thus help the firm save agency costs. Our theory, in contrast to the latter two papers, is based on standard preference and full rationality of the agent. Moreover, all papers 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.

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 **?**'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 whereas **?** considers contracting under full commitment, we assume the principal lacks commitment power. 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 **?**. He analyzes a situation where a principal faces an agent whose output depends on both his exogenous type and effort. He 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 lowering 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 by using a quota contract that induces incentive-system gaming by the agent. This endogenous information asymmetry in turn increases her commitment ability.

Finally, our dynamic moral hazard model features a 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, **?** and **?**. Unlike our model, these papers do not allow gaming activities by the agent: once produced, all outputs are immediately available the principal.

The paper is organized as follows. The model is set up in Section 2. Section 3 analyzes the benchmark case of full contractual commitment by the principal. We then study renegotiation-proof contracts in Section 4. Section 5 concludes. Longer proofs are relegated to the appendix.

2 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 round of production, the agent chooses $e \in \{0, 1\}$. The choice of e is unobservable by the principal, but the cost of effort in period t, denoted by $c_t(e) = c_t e$, is commonly known.

The agent's effort exerted in each round of production stochastically determines the output $y \in \{0, 1\}$ for that round. Let $p_e \in (0, 1)$ denote the probability of output $y = 1$ given effort e, where $p_0 \equiv p_1 - \Delta$ for some $\Delta \in (0, p_1)$. We say the agent puts in effort in a certain round if and only if $e = 1$. The effort choice and output of each round are independent. We use y_t to denote the total output in period t. Thus, $y_1 \in \{0, 1, 2\}$, and $y_2 \in \{0, 1\}$. The agent has full knowledge of the his own effort choice and outputs once they are realized. The values of each unit of output for the principal and the agent are normalized to one and zero respectively. The outside option of each player is zero.

Our model of dynamic moral hazard allows the agent to game the incentive system. Specifically, we assume the principal observes neither the agent's total output at the end of each period, nor the round at which the output is produced within the first period. Moreover, outputs produced in different rounds or periods are contractually and observationally indistinguishable to the principal. Furthermore, it is feasible for the agent to privately store outputs produced in period 1 and turn them in at the end of the second period. As a result, it is possible for the agent to game the incentive system by "pushing out" some first-period output to the second period. There is a stage of reporting output at the end of each period. The reported output is verifiable and contractible. That is, at the end of the first period, the agent can report $r_1 \in \{0, 1, 2\}$ if and only if $r_1 \leq y_1$. If the agent reports $r_1 < y_1$, he can hide and store $y_1 - r_1$ to the second period. In $t = 2$, the maximum report is $y_2 + y_1 - r_1$.

The contracting game is as follows. At the beginning of the agency relationship, the principal offers a long-term contract, which the agent can either accept or reject. 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. The first period ends after report r_1 is made and bonus $b_1(r_1)$ is collected by the agent. At the beginning of the second period, 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 $\tilde{b}_2(r_2)$ 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, which implies the agent always selects a plan from the menu.

To summarize, the timing of the game is as follows:

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

We impose a number of assumptions on the parameters and discuss their role below.

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.*

The first assumption ensures that it is never optimal for the principal to delay bonus payments.[5)] The second assumption ensures that the 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.

To simplify the subsequent analysis and illustrate our main result in the simplest possible setting, we impose the following assumptions on the costs of effort:

5)See **?**, for example.

Assumption 2 *(1)* $c_1 < \bar{c}_1 \equiv \frac{(1-p_1)\Delta^2}{p_1(1-\Delta)}$*; and* $c_2 < \bar{c}_2 \equiv \frac{\Delta^2}{p_1}$*. (2)* $c_1 \leq c_2$.

The first assumption implies that if the agent is myopic (i.e., $\rho = 0$), the following linear contract is optimal: $b_1(r_1) = r_1 \frac{c_1}{\Delta}$, and $b_2(r_2; r_1) = r_2 \frac{c_2}{\Delta}$. The second assumption states that the cost of effort is increasing over time. This is consistent with the standard assumption of increasing marginal effort cost. For example, it is likely that a salesperson has to exert a higher effort to generate later sales than earlier ones, because he has depleted "easier" sales.

Whereas the linear contract is optimal if the agent is myopic, it is not necessary the case if he is sufficiently patient. Intuitively, the linear contract motivates the agent only by providing immediate rewards, but the necessary rewards can be lowered if the contract also contain dynamic incentives. Specifically, the principal can punish the agent in $t = 2$ if he performs poorly in $t = 1$. This intuition is formalized in Section 3 where we consider a benchmark setting in which the principal can fully commit to a long-term contract. The punishment is, however, costly to the principal as valuable agent effort is lost. Therefore, a potential problem of such contract is that at the beginning of $t = 2$, the principal may be tempted to renegotiate with the agent, rendering the threat of punishment incredible. We then show in Section 4 that, by offering a quota-based contract, the principal can deliberately raise the implicit cost of renegotiation, and hence mitigate this commitment problem.

3 Benchmark Case: Full Commitment to Long-term Contract

In this section, we assume the principal is able to fully commit to the long-term contract proposed at the beginning of the agency relationship. Observe first that it is without loss to focus on contracts that induce truthful reporting of first-period output y_1 at the end of $t = 1$. The reason is that for any contract that induces agent withholding some first-period output, one can construct a corresponding contract that induces the same effort profile but with truthful reporting in the first period. As $\rho < \delta < 1$, the modified contract is more profitable to the principal.

Suppose the principal would like the agent to exert effort in every contingency (i.e., in each round of production regardless of history). Let's work out the optimal contract by backward induction. A minimum bonus of $\frac{c_2}{\Delta}$ for the second-period output is necessary for soliciting effort in $t = 2$, which implies an agency rent of $(p_1 - \Delta) c_2 / \Delta$.[6)] Moving backward to the first period, the agent's continuation payoff does not vary with his first-period output, and the incentives for effort must be provided only through immediate rewards. To solicit effort for the first period, a unit bonus of $b_1 = \frac{c_1}{\Delta}$ is required. On

[6)] Given bonuses of $b_2(1; r_1) = b_2$ and $b_2(0; r_1) = 0$, the agent is willing to put in effort in $t = 2$ if and only if $p_1 b_2 - c_2 \geq p_0 b_2$. Hence, $b_2(1; r_1) = \frac{c_2}{\Delta}$, and the resulting agent's payoff is $p_1 b_2(1; r_2) - c_2 = \frac{(p_1 - \Delta) c_2}{\Delta}$.

the other hand, to ensure the agent is willing to report his output truthfully, it is necessary that the unit bonus b_1 satisfies $b_1 + \rho(p_1 - \Delta) c_2/\Delta \geq \rho c_2/\Delta$, or $b_1 \geq \rho(1 - (p_1 - \Delta)) c_2/\Delta$. To summarize,

Lemma 1 *In the optimal linear contract,* $b_1(r_1) = r_1 \frac{1}{\Delta} \max\{c_1, \rho c_2(1 - (p_1 - \Delta))\}$ *for all* $r_1 \in \{0, 1, 2\}$ *and* $b_2(r_2; r_1) = r_2 \frac{c_2}{\Delta}$ *for all* $r_2 \in \{0, 1\}$*.*

The principal can save some agency costs for inducing first-period effort if she does NOT solicit effort from the agent following a poor first-period performance. By "firing" the agent at the end of $t = 1$, the agent is stripped of the second-period agency rent. Facing the termination threat, he is willing to put in effort in the first period even if the immediate bonus is lowered. Clearly, the principal suffers as she loses the agent's effort in the second period in some contingencies. Whether the principal finds it optimal to make use of the dynamic incentives depends on the trade-off between saving agency cost in $t = 1$ and the loss in valuable effort in $t = 2$.

A particularly important class of contracts is one that induces no effort following $y_1 = 0$. Throughout the paper, we call such a contract a **termination contract**. Formally,

Definition 1 *A contract is a termination contract if it induces effort following all histories except* $y_1 = 0$*.*

The following lemma characterizes the optimal termination contract:

Lemma 2 *In the optimal termination contract,* $b_2(0; r_1) = 0$ *for all* $r_1 \in \{0, 1, 2\}$*,* $b_2(1; 0) = 0$*,* $b_2(1; r_2) = \frac{c_2}{\Delta}$ *for* $r_1 \in \{1, 2\}$*. Moreover,* $b_1(0) = 0$*,*

$$b_1(1) = \max\left\{\frac{c_1 - \rho(p_1 - \Delta) c_2}{\Delta}, 0\right\}, \textit{ and}$$

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

Proof. In the Appendix ■

The following proposition simplifies the search for the optimal long-term contract:

Proposition 1 *Suppose the ratios of effort costs satisfy* $\frac{c_2}{c_1} \in \left[1, \frac{1}{\rho} \min\left\{\frac{1}{p_1 - \Delta}, \frac{1}{1 - (p_1 - \Delta)}\right\}\right]$*. Then the optimal long-term contract is either the optimal linear contract, or the optimal termination contract.*

The assumption of the theorem requires that the effort cost does not increase too rapidly over time. Specifically, it ensures that (i) all bonus payments in the optimal contract are positive; and (ii) the constraint for truthful output reporting is not binding.[7] The theorem states that in this region of costs, there

[7]To see (ii), note that the incentive compatibility constraint for truthful reporting is not binding whenever $\max\left\{\frac{c_1}{\Delta}, \rho \frac{(1 - (p_1 - \Delta)) c_2}{\Delta}\right\} = \frac{c_1}{\Delta}$.

are only two candidates for the optimal long-term contract: a linear contract and a termination contract. This cost region is particularly interesting for our investigation because it implies that under full contractual commitment, a quota-based contract does not arise (condition (i) above), and incentive-system gaming can be deterred at no extra cost to the principal (condition (ii) above). In the rest of the paper, we will focus on this parameter region and show in the subsequent section that if the principal cannot refrain from renegotiation, a quota-based contract that induces incentive-system gaming may emerge as the optimal contract.

Assumption 3 *The ratio of effort costs satisfy* $\frac{c_2}{c_1} \in \left[1, \frac{1}{\rho} \min\left\{\frac{1}{p_1-\Delta}, \frac{1}{1-(p_1-\Delta)}\right\}\right]$.

Using Proposition 1, it suffices to compare the optimal linear contract with the optimal termination contract. 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,

$$C \equiv \underbrace{(1-p_1)^2}_{\Pr(y_1=0)} \times \delta \times \underbrace{\left(p_1\left(1-\frac{c_2}{\Delta}\right)-(p_1-\Delta)\right)}_{\text{principal's loss in } t=2 \text{ when } y_1=0.}; \text{ and}$$

$$S \equiv \underbrace{p_1(2-p_1)}_{\Pr(y_1>0)} \times \rho \times \underbrace{\left(\frac{(p_1-\Delta)\, c_2}{\Delta}\right)}_{\text{agent's rent in } t=2 \text{ when } y_1\neq 0}. \tag{1}$$

The principal adopts the optimal termination contract if the saving in agency costs S exceeds the loss in valuable effort C. The optimal termination contract is favored if c_2 is large. Intuitively, 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. 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.

On the other hand, the optimal termination contract is also favored if p_1 is large. Intuitively, a large value of p_1 means it is unlikely that the agent fails both rounds in $t = 1$, i.e., $\Pr(y_1 > 0)$ is a lot higher than $\Pr(y_1 = 0)$. Moreover, from the agent's perspective, a large value of p_1 implies a larger agency rent $(p_1 - \Delta)\, c_2/\Delta$ in $t = 2$. These observations are formalized in the following corollary.

Corollary 1 *(i) For each c_1 and c_2, there exists a $p_1^{FC}(c_2)$ such that the optimal termination contract is the optimal long-term contract if and only if $p_1 \geq p_1^{FC}(c_2)$.*

(ii) For each c_1 and p_1, there exists a $c_2^{FC}(p_1)$ such that the optimal termination contract is the optimal long-term contract if and only if $c_2 \geq c_2^{FC}(p_1)$.

The proof of the corollary is omitted as it follows from the straightforward observations that the term $S - C$ is increasing in both p_1 and c_2, and is positive at $p_1 = 1$ and $c_2 = \bar{c}_2$. According to the corollary, neither $p_1^{FC}(c_2)$ nor $c_2^{FC}(p_1)$ depends on c_1. This is because neither the losses in valuable effort C nor savings in agency cost S depends on the absolute size of c_1.

4 Renegotiation-Proof Contract

Now we move to the main analysis in which the principal may renegotiate with the agent at the beginning of $t = 2$. By the renegotiation-proofness principle (Dewatripont (1989)), 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. Whereas the optimal linear contract is renegotiation-proof (as it maximizes efficiency), the optimal termination contract is not (as it involves inefficient termination). As a result, with a termination contract, it is in the interest of both parties to renegotiate following $y_1 = 0$. The key message of our analysis below is that a quota-based contract can help the principal mitigate the commitment problem while making effective use of dynamic incentives like the termination contract.

We first introduce a class of contracts that is key for our analysis:

Definition 2 *A contract is a quota contract if for some $B, \beta \in \mathbb{R}_+$,*

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

Under a quota contract, the agent is paid in the first period if and only if he turns in two units of output, i.e., the "quota" of the first period is two. The bonus for meeting the quota is B. The payment in the second period depends on whether the first-period quota is met or not. If the agent meets the quota, his second-period quota is one, and the associated bonus is c_2/Δ. Otherwise, his second-period quota is two, and the associated bonus is β. By choosing β and B appropriately, the quota contract can induce the same effort profile as the termination contract, i.e., the agent is induced to exert effort in both rounds in $t = 1$, and in $t = 2$ if and only if $y_1 = 1, 2$. Below, we identify conditions under which the quota contract induces this effort profile and is immune to renegotiation.

The use of a quota contract has the following implications. First, different from the termination contract, an agent with $y_1 = 1$ is "encouraged" to game the incentive system, i.e., he would like to carry his first-period output to $t = 2$. 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. Second, the agent with $y_1 = 0$ would not receive any bonus, irrespective of his performance in $t = 2$. Thus, the quota contract is necessarily inefficient, as the agent would not put in effort in $t = 2$ whenever $y_1 = 0$.

As a result of the last implication, the principal may be tempted to renegotiate with the agent at the beginning of $t = 2$, with the goal of soliciting effort by the agent with $y_1 = 0$. The *gains from renegotiation* can thus be written as

$$\underbrace{\frac{(1-p_1)^2}{2p_1(1-p_1)+(1-p_1)^2}}_{\Pr(y_1=0|r_1=0)} \times \underbrace{\left(p_1\left(1-\frac{c_2}{\Delta}\right)-(p_1-\Delta)\right)}_{\text{principal's loss in } t=2 \text{ when } y_1=0.}.$$

Improving the contract for the agent with $y_1 = 0$, however, involves an indirect cost: extra information rent must be offered to the agent with $y_1 = 1$, as his first-period output y_1 is his private information. More precisely, let $\left\{b_2^{y_1}(r_2)\right\}_{y_1=0,1}$ be the menu of contracts offered at the renegotiation stage. If the principal decides to induce effort by the agent with $y_1 = 0$, the minimum bonus for his second-period success is $b_2^0(1) = c_2/\Delta$. At the same time, to prevent the agent with $y_1 = 1$ from shirking in $t = 2$, i.e., misrepresenting his type as $y_1 = 0$, bonus $b_2^1(2)$ has to be set sufficiently high: $p_1 b_2^1(2) - c_2 \geq c_2/\Delta$, or $b_2^1(2) \geq \frac{(1+\Delta)c_2}{\Delta p_1}$. This extra information rent for the agent with $y_1 = 1$ is referred to as the *costs of renegotiation*, which can be written as

$$\underbrace{\frac{2p_1^2(1-p_1)}{2p_1(1-p_1)+(1-p_1)^2}}_{\Pr(y_1=1|r_1=0)} \times \underbrace{p_1\left(\frac{(1+\Delta)c_2}{\Delta p_1}-\beta\right)}_{\text{info rent for } y_1=1.}.$$

The quota contract is renegotiation-proof if the costs of renegotiation outweigh the gains from renegotiation. It is straightforward to show that this renegotiation-proofness constraint can be translated into the requirement that β is sufficiently small.

Lemma 3 *A quota contract that induces the effort profile of the termination contract is renegotiation-proof if and only if*

$$\beta \leq \left(\frac{1}{2}+\frac{1}{2p_1}+\Delta\right)\frac{c_2}{p_1\Delta} - \frac{1-p_1}{2p_1^3}\Delta. \tag{2}$$

Our goal is to look for the optimal renegotiation-proof quota contract that induces the effort profile of the termination contract. The lemma below states the condition that ensures its existence and provides a full characterization:

Lemma 4 *Suppose*

$$\underbrace{\frac{2p_1^2(1-p_1)}{2p_1(1-p_1)+(1-p_1)^2}p_1\left(\frac{\rho c_2 - c_1}{p_1\rho\Delta}\right)}_{\textit{Costs of Renegotiation}} \geq \underbrace{\frac{(1-p_1)^2}{2p_1(1-p_1)+(1-p_1)^2}\left(\Delta - p_1 c_2/\Delta\right)}_{\textit{Gains from Renegotiation}}. \tag{3}$$

The optimal renegotiation-proof quota contract that induces the effort profile of the termination contract has $\beta = \beta^*$ and $B = B^*$ defined as follows:

$$\beta^* \equiv \frac{\rho \Delta c_2 + c_1}{\rho p_1 \Delta} \text{ and } B^* \equiv \begin{cases} B^E \equiv \frac{2c_1 - \rho(p_1 - \Delta)c_2}{\Delta} & \text{if } p_1 \geq \hat{p}_1 \equiv \frac{\rho c_2 \Delta + c_1}{2c_1} \\ B^T \equiv \frac{\rho(\Delta - (p_1 - \Delta)p_1)c_2 + c_1}{p_1 \Delta} & \text{if } p_1 < \hat{p}_1 \end{cases}. \quad (4)$$

Proof. In the Appendix. ■

The optimal quota contract identified in Lemma 4 has a couple of noteworthy features. First, the reward for achieving the quota of 2, denoted by B^*, exceeds $\frac{c_1}{\Delta}$, the "slope" of the optimal linear contract. In this sense, we interpret the contract features a discrete jump for achieving the quota. The reason for the existence of the discrete jump is as follows. The agent who marginally fails to meet the quota is rewarded ONLY by the second-period agency rent, whereas the agent who manages to meet the quota is rewarded by both a current bonus and the second-period agency rent. However, the second-period agency rent offered in the former case must exceed that of the latter case, so that sufficient work incentives are provided for the agent to put in effort in the second round of the first period following a first-round failure.

Second, the quota in the second period depends on the reported output in the first period. Specifically, if $r_1 < 2$, the agent receives a bonus in period 2 if and only if $r_2 = 2$; otherwise, if $r_1 = 2$, the agent receives a bonus if and only if $r_2 = 1$. Thus, the agent faces a more challenging quota in the second period if his first-period performance is below par. This contractual arrangement is observed in practice. In their empirical study for the incentive structure and work behavior of mortgage loan officers, **?** reports that in a typical contract in the industry, "any deficit in a monthly quota is carried over to the subsequent month, thus augmenting that month's quota." Finally, it is worth noting that this feature of the contract is inconsistent with the ratchet effect as in **?**, which predicts that a better early performance is followed by a tougher contract.

To conclude this subsection, we discuss conditions under which inequality (3) in Lemma 4, which ensures the existence of a desired renegotiation-proof contract, is likely to hold. Observe first that as the gains from renegotiation (the right-hand side of inequality (3)) is positive, a necessary condition for (3) to hold is that the costs of renegotiation (the left-hand side of inequality (3)) is positive, or equivalently, $\frac{c_2}{c_1} > \frac{1}{\rho}$. Now suppose p_1 is high. Following a report of $r_1 = 0$ in the first period, the principal believes that it is likely that the agent has $y_1 = 1$, i.e., $\Pr(y_1 = 1|r_1 = 0)$ far exceeds $\Pr(y_1 = 1|r_1 = 0)$. Therefore, renegotiation is likely to be unprofitable for the principal.

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

Corollary 2 *Suppose* $\frac{c_2}{c_1} > \frac{1}{\rho}$.

(i) For each c_1, c_2, *and* Δ, *there exists a* $p_1^{RP} < 1$ *such that inequality (3) holds if and only if* $p_1 \geq p_1^{RP}$.

(ii) For each c_1, p_1, *and* Δ, *there exists a* $c_2^{RP} < \bar{c}_2$ *such that inequality (3) holds if and only if* $c_2 \geq c_2^{RP}$.

Proof. In the Appendix. ■

4.1 Quota Contract vs Linear Contract

In this subsection, we compare the performance of the optimal quota contract and the optimal linear contract under the renegotiation-proofness constraint. We identify the cost to the principal imposed by the renegotiation-proofness constraint, as well as sufficient conditions under which the optimal quota contract outperforms the optimal linear contract. For the remainder of this subsection, we assume $S - C > 0$ so that if contract commitment is perfect, the optimal termination contract is preferred.

Suppose renegotiation is possible and inequality (2) holds. Using the optimal quota contract identified in Lemma 4, the principal can implement the same effort profile as the termination contract, albeit at a higher cost. The *losses from the lack of commitment,* denoted by L, is defined as the difference in the principal's payoffs under the optimal termination contract and optimal quota contract. We can express the condition for optimal quota contract outperforming the optimal linear contract as follows:

$$\underbrace{S}_{\text{Saving in Agency Cost}} - \underbrace{C}_{\text{Loss in Effort}} > \underbrace{L}_{\text{Losses from Lack of Commitment}}. \tag{5}$$

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 (5) holds, we can decompose the loss term L into three components: (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, $L = L_1 + L_2 + L_3$, where

$$L_1 \equiv \underbrace{p_1^2\left(\max\left\{\frac{\rho\Delta c_2 + (1-2p_1)\,c_1}{p_1\Delta}, 0\right\}\right)}_{\text{cost of inducing truthful report in } t=1};$$

$$L_2 \equiv \underbrace{2p_1\,(1-p_1)\,(1-\delta)}_{\text{cost of system gaming}}; \text{ and}$$

$$L_3 \equiv \underbrace{2p_1\,(1-p_1)\left((\delta-\rho)\,\frac{c_1 - \rho\,(p_1-\Delta)\,c_2}{\rho\Delta}\right)}_{\text{cost of delaying payment}}.$$

First, the cost of inducing truthful report L_1, arises from preventing the agent with $y_1 = 2$ from withholding the produced outputs to $t = 2$. Observe that it is zero if p_1 exceeds $\frac{1}{2}$ and c_1 is close to c_2. In this case, the principal has to pay a large bonus B just for inducing effort in $t = 1$, so the constraint on truthful reporting is not binding. Second, the cost of system gaming L_2, arises because if the agent has $y_1 = 1$, he delays reporting it to $t = 2$, imposing a time-discounting loss to the principal. Finally, the cost of delaying payment L_3, arises from the difference in time preference between the principal and the agent. Specifically, if $y_1 = 1$, the agent is paid only in $t = 2$, and the principal suffers as she discounts future payoffs less than the agent.

This decomposition allows us to identify conditions that favor the quota contract. First, if ρ is getting close to δ, the cost of delaying payment L_3 decreases. Moreover, as future incentives is more effective for inducing effort, the term $S - C$ increases, and the quota contract is more appealing. Second, the overall effect of an increase in δ is ambiguous: it lowers the cost of system gaming L_2, but increases the loss in effort C, as well as the cost of delaying payment L_3. Third, the quota contract dominates the linear contract if p_1 is sufficiently large. In this case, all loss terms become small, whereas the saving in agency cost remains strictly positive. Finally, the quota contract is favored if the degree of moral hazard, as measured by ratios $\frac{c_1}{\Delta}$ and $\frac{c_2}{\Delta}$, increases. Specifically, if the quota contract dominates the linear contract for some c_i' and Δ', then it does so for all $c_i'' > c_i'$ and $\Delta'' < \Delta'$. The reason is that the term $S-C$ is increasing in $\frac{c_i}{\Delta}$, and that $S - C - L$ is linear in c_i and decreasing in Δ.

The following proposition summarizes the discussion above by stating the sufficient conditions for the superiority of the optimal quota contract over the optimal linear contract.

Proposition 2 *Suppose $p_1 > \frac{1}{2}$ and $\Delta < p_1\frac{p_1+1}{p_1+2}$. There exists a $\bar{\rho} \in (0,1)$ and a $\hat{c}_2 : [\bar{\rho}, 1] \times [0, \bar{c}_1] \to [0, \bar{c}_2]$ such that for all $\rho \geq \bar{\rho}$ and $c_2 > \hat{c}_2(\rho, c_1)$, the optimal quota contract is renegotiation-proof and more profitable than the linear contract.*

Proof. In the Appendix ■

4.2 Optimality of Quota Contract

In this subsection, we provide sufficient conditions that ensure the optimal quota contract is the optimal renegotiation-proof contract. We have seen above that a benefit of using the quota contract is that by inducing incentive-system gaming by the agent with mediocre performance, it helps the principal commit to credibly punish the agent with poor performance. One may wonder that other contracts may achieve the same goal. Consider, for example, a contract that induces the agent with $y_1 = 2$ to withhold all outputs and the agent with $y_1 = 1$ to report truthfully at the end of $t = 1$. A problem of this contract is that it leaves a high agency rent for the agent with $y_1 = 2$ at the beginning of the second period, making it profitable for the principal to offer a new contract that solicits the effort by the agent with $y_1 = 0$ (and that is rejected by the agent with $y_1 = 2$). Alternatively, the principal may offer a contract that induces the agent with any positive y_1 to withhold all outputs to the second period. However, compared with the optimal quota contract, this contract incurs a high agency cost, as wage payments and receipt of outputs are delayed, whereas the agent must receive the same agency rent.

The principal can consider contracts that implement other effort profiles. However, Proposition 1 implies that the benefit of doing so is limited, especially if the losses from the lack of commitment L (defined in the previous subsection) is small. These observations lead us to conclude that if the optimal quota contract outperforms the optimal linear contract, it is likely to be the optimal renegotiation-proof contract. Formally,

Proposition 3 *Suppose* $p_1 > \frac{1}{2}$ *and* $\Delta < \frac{1}{2}\min\left\{p_1, p_1 + 3 - \sqrt{p_1^2 - 6p_1 + 13}\right\}$*. There exists a* $\bar{\rho} \in (0,1)$ *and a* $\hat{c}_2 : [\bar{\rho}, 1] \times [0, \bar{c}_1] \to [0, \bar{c}_2]$ *such that for all* $\rho \geq \bar{\rho}$ *and* $c_2 > \hat{c}_2(\rho, c_1)$*, the optimal quota contract is the optimal renegotiation-proof contract.*

Proof. In the Appendix ■

5 Concluding Remarks

Our model sheds light on the widespread use of the quota contract by identifying a novel benefit of incentive-system gaming. Below, we discuss some alternative specifications of our model. First, we allow 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).[8)] This specification makes the quota contract renegotiation-proof for a wider range of parameters, and thus strengthens the result that the quota contract can outperform the linear contract.

[8)]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 $\tilde{b}_2(r_2; r_1) \geq b_2(r_2; r_1)$.

Second, the monitoring technology assumed allows the agent to hide produced outputs from the principal, store and report them to the principal only in a later period. If we instead assume the principal can directly monitor the agent's production, so that any produced output is immediately known to the principal, the principal is always weakly better off if she has full contractual commitment power (Holmstrom (1979)). However, it is not necessarily the case if the principal cannot refrain from renegotiation at the beginning of period 2. With this direct monitoring technology, the principal cannot commit to a termination contract, making the use of future incentives infeasible. As a result, the principal may intentionally adopt a worse monitoring technology to enforce a contract that involves inefficient punishment.[9)]

There are a couple of 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 believe that our main result will still hold: a quota contract can outperform linear contracts, particularly when the moral hazard problem is severe.

Another natural extension is 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 agent's "pushing out", and 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.

6 Appendix

Proof of Lemma 2 It is straightforward to see that $b_2(1;2) = b_2(1;1) = c_2/\Delta$, and $b_1(0) = b_2(1;0) = b_2(0;r_1) = 0$ for all $r_1 \in \{0,1,2\}$. It remains to specify $b_1(1)$ and $b_1(2)$. The agent exerts effort in the second round in $t = 1$ following a first-round failure if and only if

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

$$b_1(1) \geq \frac{c_1 - \rho(p_1 - \Delta)c_2}{\Delta}. \tag{6}$$

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

$$p_1 b_1(2) + (1-p_1)b_1(1) + \rho\left(\frac{p_0 c_2}{\Delta}\right) - c_1 \geq p_0 b_1(2) + (1-p_0)b_1(1) + \rho\left(\frac{p_0 c_2}{\Delta}\right),$$

9) This finding is reminiscent of Cremer (1995).

which can be simplified to

$$b_1(2) \geq \frac{c_1}{\Delta} + b_1(1). \tag{7}$$

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

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

which can be simplified to

$$p_1 b_1(2) + (1-2p_1) b_1(1) \geq \frac{c_1 - (1-p_1)\rho(p_1-\Delta) c_2}{\Delta}. \tag{8}$$

Finally, we need to ensure that the agent with $y_1 = 2$ is willing to report truthfully, rather than reporting only one unit and shirking in the second period. This requires

$$b_1(2) + \rho\frac{(p_1-\Delta) c_2}{\Delta} \geq b_1(1) + \rho\frac{c_2}{\Delta} \Leftrightarrow b_1(2) \geq b_1(1) + \rho\frac{(1-(p_1-\Delta)) c_2}{\Delta}. \tag{9}$$

It is straightforward algebra to show that the smallest values of $b_1(1)$ and $b_1(2)$ that satisfy inequalities (6), (7), (8), and (9) are those stated in the Lemma statement. Q.E.D.

Proof of Theorem 1 We only need to consider contracts that induce one of the following five effort profiles:

(i) effort is exerted following every history;

(ii) effort is exerted following every history except in $t = 2$ following $y_1 = 0$;

(iii) effort is exerted following every history except in $t = 2$ following $y_1 \in \{0, 1\}$;

(iv) effort is exerted following every history except in the second round of $t = 1$ following a first-round failure, as well as in $t = 2$ following $y_1 = 0$;

(v) effort is exerted if and only if production has always been successful.

Effort profile (i): As explained at the beginning of Section 3, the optimal linear contract minimizes the agency cost necessary to implement this effort profile. Using Assumption 3, the principal's profit is

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

Effort profile (ii): As shown in Lemma 2, the optimal termination contract minimizes the agency cost necessary to implement this effort profile. Using

Assumption 3, the principal's profit is

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

Effort profile (iii): It is straightforward to see that $b_2\left(1;r_1\right)=0$ for $r_1\in\{0,1\}$, $b_2\left(0;2\right)=0$, and $b_2\left(1;2\right)=\frac{c_2}{\Delta}$. Moreover, the agent is willing to put in effort in $t=1$ following a first-round failure if and only if $b_1\left(1\right)\geq\frac{c_1}{\Delta}$. On the other hand, the agent is willing to put in effort in $t=1$ following a first-round success if and only if

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

It is straightforward to show that the two inequalities above imply the agent is willing to put in effort in the first round of $t=1$. Moreover, the agent is always willing to report truthfully. The lowest values of wages that satisfy both inequalities are $b_1\left(1\right)=\frac{c_1}{\Delta}$, and $b_1\left(2\right)=\max\left\{\frac{2c_1-\rho\left(p_1-\Delta\right)c_2}{\Delta},0\right\}$. The assumption that $\frac{c_2}{c_1}\leq\frac{1}{\rho\left(p_1-\Delta\right)}$ implies $b_1\left(2\right)=\frac{2c_1-\rho\left(p_1-\Delta\right)c_2}{\Delta}$. The principal's profit under this contract is thus

$$\Pi^{(iii)}\equiv p_1^2\left(2-\frac{2c_1-\rho\left(p_1-\Delta\right)c_2}{\Delta}\right)+2p_1\left(1-p_1\right)\left(1-\frac{c_1}{\Delta}\right)+\delta\left(p_1^3\left(1-\frac{c_2}{\Delta}\right)+\left(1-p_1^2\right)\left(p_1-\Delta\right)\right).$$

Effort profile (iv): It is straightforward that $b_2\left(1;r_1\right)=\frac{c_2}{\Delta}$ for $r_1\in\{1,2\}$, and $b_2\left(1;0\right)=0$. Moreover, as effort is not induced following first-round failure, we have $b_1\left(1\right)=0$.

The agent is willing to put in effort following a first-round success in $t=1$ if and only if

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

The agent is willing to put in effort in the first round of $t = 1$ if and only if

$$
\begin{aligned}
& p\left(pb_1(2) - c_1 + \rho\frac{(p_1-\Delta)c_2}{\Delta}\right) + (1-p)\rho(p_1-\Delta)\frac{(p_1-\Delta)c_2}{\Delta} - c_1 \\
& \geq (p_1-\Delta)\left(pb_1(2) - c_1 + \rho\frac{(p_1-\Delta)c_2}{\Delta}\right) + (1-(p_1-\Delta))\rho(p_1-\Delta)\frac{(p_1-\Delta)c_2}{\Delta} \\
& \Leftrightarrow p\left(pb_1(2) - c_1 + \rho\frac{(1-(p_1-\Delta))(p_1-\Delta)c_2}{\Delta}\right) - c_1 \\
& \geq (p_1-\Delta)\left(pb_1(2) - c_1 + \rho(1-(p_1-\Delta))\frac{(p_1-\Delta)c_2}{\Delta}\right) \\
& \Leftrightarrow b_1(2) \geq \frac{1}{p\Delta}\left((\Delta+1)c_1 - \rho(1-(p_1-\Delta))(p_1-\Delta)c_2\right).
\end{aligned}
$$

The agent is willing to report truthfully if and only if

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

The lowest value that satisfies all inequalities above is

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

The assumption that $\frac{c_2}{c_1} \leq \frac{1}{\rho(p_1-\Delta)}$ implies $b_1(2) = \frac{1}{\Delta p}\left((\Delta+1)c_1 - \rho(1-(p_1-\Delta))(p_1-\Delta)c_2\right)$. The principal's profit under this contract is thus

$$
\begin{aligned}
\Pi^{(iv)} \equiv\, & p_1^2\left(2 - \frac{(1+\Delta)c_1 - \rho(p_1-\Delta)(1-(p_1-\Delta))c_2}{\Delta p_1}\right) + (1-p_1)(p_1+(p_1-\Delta)) \\
& + \delta\left((1-(1-p_1)(1-(p_1-\Delta)))p_1\left(1-\frac{c_2}{\Delta}\right) + (1-p_1)(1-(p_1-\Delta))(p_1-\Delta)\right).
\end{aligned}
$$

Effort profile (v): It is straightforward to see that $b_2(1; r_1) = 0$ for $r_1 \in \{0, 1\}$, and $b_2(1; 2) = \frac{c_2}{\Delta}$. Moreover, $b_1(1) = b_1(0) = 0$. The agent exerts effort in the second round in $t = 1$ following a success if and only if

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

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

$$
\begin{aligned}
& \Delta\left(p_1\left(b_1(2) + \rho\left(\frac{(p_1-\Delta)c_2}{\Delta}\right)\right) - c_1\right) \geq c_1 \\
& \Leftrightarrow b_1(2) \geq \frac{(1+\Delta)c_1 - \rho p_1(p_1-\Delta)c_2}{p_1\Delta}.
\end{aligned}
$$

Clearly, the lowest value of $b_1(2)$ is $\max\left\{0, \frac{(1+\Delta)c_1-\rho p_1(p_1-\Delta)c_2}{p_1\Delta}\right\}$. The assumption that $\frac{c_2}{c_1} \leq \frac{1}{\rho(p_1-\Delta)}$ implies $b_1(2) = \frac{(1+\Delta)c_1-\rho p_1(p_1-\Delta)c_2}{p_1\Delta}$. The principal's profit is thus

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

Below, we show that the inducing effort profile (ii) gives the principal a higher profit than inducing effort profiles (iii)-(v). To see $\Pi^T > \Pi^{(iii)}$,

$$\begin{aligned}\Pi^T - \Pi^{(iii)} &= 2\frac{p_1(1-p_1)}{\Delta}\left(\Delta^2\delta - (\delta p_1 - \rho(p_1-\Delta))c_2\right)\\ &> 2\frac{p_1(1-p_1)}{\Delta}\left(\Delta^2\delta - (\delta p_1 - \rho(p_1-\Delta))\frac{\Delta^2}{p_1}\right)\\ &= 2p_1(1-p_1)\Delta(\delta(1-\delta) + \rho(p_1-\Delta)\Delta) > 0.\end{aligned}$$

The first inequality follows from $c_2 \leq \bar{c}_2$. Next, to see $\Pi^T > \Pi^{(iv)}$,

$$\begin{aligned}\Pi^T - \Pi^{(iv)} &= \frac{1}{\Delta}\left[\Delta^2(1-p_1)(1+\delta\Delta) - p_1c_1(1-\Delta) + p_1c_2(-\Delta\delta(1-p_1) + (p_1-\Delta)\rho(1-\Delta))\right]\\ &\geq \frac{1}{\Delta}\left[\Delta^3(1-p_1)\delta + p_1c_2(-\Delta\delta(1-p_1) + (p_1-\Delta)\rho(1-\Delta))\right]\\ &= \frac{p_1}{\Delta}\left[\delta(1-p_1)\Delta\left(\frac{\Delta^2}{p_1} - c_2\right) + (p_1-\Delta)(1-\Delta)p_1c_2\right] > 0.\end{aligned}$$

The first inequality follows from $c_1 \leq \bar{c}_1$ and the last inequality follows from $c_2 \leq \bar{c}_2$. Finally, to see $\Pi^T > \Pi^{(v)}$

$$\begin{aligned}\Pi^T - \Pi^{(v)} &= \frac{1}{\Delta}\left(\Delta^2(1-p_1)(1+2p_1\delta) - p_1(1-\Delta)c_1 - 2p_1c_2(1-p_1)(\delta p_1 - \rho(p_1-\Delta))\right)\\ &\geq \frac{1}{\Delta}\left(\Delta^2(1-p_1)(1+2p_1\delta) - (1-p_1)\Delta^2 - 2\Delta^2(1-p_1)(\delta p_1 - \rho(p_1-\Delta))\right)\\ &= 2\Delta(1-p_1)\rho(p_1-\Delta) > 0\end{aligned}$$

The first inequality follows from $c_1 \leq \bar{c}_1$ and $c_2 \leq \bar{c}_2$. Q.E.D.

Proof of Lemma 3

Consider the principal's problem at the beginning of $t = 2$ following a report $r_1 = 0$. If the principal does not renegotiate, her expected payoff is

$$\Pi_2^O \equiv \frac{2p_1(1-p_1)}{2p_1(1-p_1) + (1-p_1)^2}(1 + p_1(1-\beta)) + \frac{(1-p_1)^2}{2p_1(1-p_1) + (1-p_1)^2}p_0.$$

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 $\left\{b_2^{y_1}(r_2)\right\}_{y_1=0,1}$ of the following form:

$$b_2^0(0;0) = 0; b_2^0(1;0) = b_2^0(2;0) = c_2/\Delta;$$
$$b_2^1(0;0) = b_2^1(1;0) = 0; b_2^1(2;0) = \tilde{b}.$$

Bonus $\tilde{b}$ has to be chosen such that (i) agent with y_1 chooses $b_2^{y_1}(\cdot)$; (ii) agent with $y_1 = 1$ exerts effort in $t = 2$; and (iii) $\tilde{b} \geq \beta$. These requirements translate into

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

Therefore, the principal's expected payoff from renegotiation is

$$\Pi_2^R \equiv \frac{2p_1(1-p_1)}{2p_1(1-p_1)+(1-p_1)^2}\left(1+p_1\left(1-\tilde{b}(\beta)\right)\right)+\frac{(1-p_1)^2}{2p_1(1-p_1)+(1-p_1)^2}p_1\left(1-\frac{c_2}{\Delta}\right).$$

Comparing Π_2^O with Π_2^R, we can conclude that for a given β, the principal find the renegotiation unprofitable if and only if

$$\underbrace{\frac{2p_1^2(1-p_1)}{2p_1(1-p_1)+(1-p_1)^2}\left(\tilde{b}(\beta)-\beta\right)}_{\text{Costs of Renegotiation}} \geq \underbrace{\frac{(1-p_1)^2}{2p_1(1-p_1)+(1-p_1)^2}(\Delta - p_1c_2/\Delta)}_{\text{Gains from Renegotiation}}. \quad (10)$$

Rearranging (10) gives (2). Q.E.D.

Proof of Lemma 4

First, we 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 exerts effort in the second round of $t = 1$ following a first-round failure if and only if

$$\beta \geq \beta^* \equiv \frac{\rho\Delta c_2 + c_1}{p_1\rho\Delta}. \quad (11)$$

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

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

Thus, for a given β, 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 $p_1\rho(p_1\beta^* - c_2) - c_1 = p_0c_1/\Delta$. 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/Δ, 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 β, we know that the 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 β, the agent with $y_1 = 2$ reports truthfully if and only if $B \geq \rho\left(\beta - \rho\frac{(p_1-\Delta)c_2}{\Delta}\right) \equiv B^T(\beta)$.

As both $B^E(\beta)$ and $B^T(\beta)$ are increasing in β, the principal would find it optimal to set β as low as possible, since doing so would not tighten the incentive constraints and renegotiation-proofness constraint (recall Lemma 3). Thus, we can focus on quota contract with $\beta = \beta^*$ and $B = B^*(\beta^*) \equiv \max\left\{B^E(\beta^*), B^T(\beta^*)\right\}$. It is straightforward algebra to show that these give rise to (4) in the lemma statement.

Finally, substituting $\beta = \beta^*$ into inequality (2) yields (3). Q.E.D.

Proof of Corollary 2

Define $RC(p_1, c_1, c_2) \equiv 2p_1(c_2 - c_1/\rho) - (1-p_1)\left(\Delta^2 - p_1c_2\right)$. Then, (**??**) 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

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

Moreover, $\lim_{p_1 \to \min\left\{1, \frac{\Delta^2}{c_2}\right\}} RC(p_1, c_1, c_2) = \left(c_2 - \frac{c_1}{\rho}\right) > 0$. Therefore, $RC(p_1, c_1, c_2) > 0$ if and only if $p_1 \geq p_1^{RP}(c_1, c_2) \equiv \frac{\left(3c_2 - 2\frac{c_1}{\rho} + \Delta^2\right) - \sqrt{\left(3c_2 - 2\frac{c_1}{\rho} + \Delta^2\right)^2 - 4c_2\Delta^2}}{2c_2}$.

Next, note that $RC(p_1, c_1, c_2)$ is increasing in c_2. Therefore, $RC(\cdot) > 0$ if and only if $c_2 > c_2^{RP}(p_1, c_1) \equiv \frac{2p_1c_1 + \Delta^2\rho(1-p_1)}{\rho p_1(3-p_1)}$. Q.E.D.

Proof of Proposition 2

As $p_1 > \frac{1}{2}$, the loss terms L_i have respective upper bounds below:

$$L_1 \le \rho p_1^2 \left(2\Delta - (2p_1 - 1)\right) \frac{c_2}{\Delta};$$
$$L_2 \le \frac{1}{2}(1-\rho);$$
$$L_3 \le \frac{1}{2}(1-\rho)\left(\frac{c_1}{\rho\Delta} - (p_1 - \Delta)\frac{c_2}{\Delta}\right).$$

Recall

$$\begin{aligned} S - C &= p_1(2-p_1)\rho\frac{(p-\Delta)c_2}{\Delta} - \delta(1-p_1)^2\left(\Delta - p\frac{c_2}{\Delta}\right) \\ &\ge p_1(2-p_1)\rho\frac{(p-\Delta)c_2}{\Delta} - (1-p_1)^2\left(\Delta - p_1\frac{c_2}{\Delta}\right). \end{aligned}$$

Define

$$\begin{aligned} \tilde{D} \equiv &\left(p_1(2-p_1)\rho\frac{(p_1-\Delta)c_2}{\Delta} - (1-p)^2\left(\Delta - p_1\frac{c_2}{\Delta}\right)\right) \\ &- \left(\rho p_1^2\left(2\Delta - (2p_1-1)\right)\frac{c_2}{\Delta} + \frac{1}{2}(1-\rho)\left(1 + \frac{c_1}{\rho\Delta} - (p_1-\Delta)\frac{c_2}{\Delta}\right)\right). \end{aligned}$$

The optimal quota contract outperforms the linear contract if $\tilde{D} > 0$. By some direct computation, we have

$$\begin{aligned} \tilde{D} = &\left[p_1\left(1 - p_1 + 2p_1^2 - 2\Delta - p_1\Delta\right) - (1-\rho)\left[p_1(2-p_1)(p_1-\Delta) - p_1^2\left(2\Delta - (2p_1-1)\right) - \frac{1}{2}(p_1-\Delta)\right]\right]\frac{c_2}{\Delta} \\ &- (1-p_1)^2\Delta - \frac{1}{2}(1-\rho)\left(\frac{c_1}{\rho\Delta} + 1\right) \\ \ge &\left[p_1\left(1 - p_1 + 2p_1^2 - 2\Delta - p_1\Delta\right) - \frac{3}{2}(1-\rho)\right]\frac{c_2}{\Delta} - (1-p_1)^2\Delta - \frac{1}{2}(1-\rho)\left(\frac{1}{\rho} + 1\right) \\ = &\left(1 - p_1 + 2p_1^2 - 2\Delta - p_1\Delta\right)\frac{p_1c_2}{\Delta} - (1-p_1)^2\Delta - \frac{1}{2}(1-\rho)\left(\frac{1}{\rho} + 1 + \frac{3c_2}{\Delta}\right) \\ \ge &\left[\left(1 - p_1 + 2p_1^2 - 2\Delta - p_1\Delta\right)\frac{p_1c_2}{\Delta} - (1-p_1)^2\Delta\right] - (1-\rho)\left(\frac{1}{\rho} + 2\right) \end{aligned}$$

where the first inequality $c_1 < \frac{(1-p_1)\Delta^2}{p_1(1-\Delta)} < \frac{\Delta}{p_1} < 2\Delta$ and that $p(2-p)(p-\Delta) - p^2(2\Delta - (2p-1)) - \frac{1}{2}(p-\Delta) < \frac{3}{2}$.[10)] The second inequality makes use of $c_2 < \frac{\Delta^2}{p_1}$.

[10)]To see that, note that the derivative of the latter term with respect to Δ is$-p^2 - 2p + \frac{1}{2} < 0$. Therefore, it achieves the minimum at $\Delta = 0$, giving $\frac{1}{2}p\left(2p + 2p^2 - 1\right) < \frac{3}{2}$.

The term $(1-\rho)\left(\frac{1}{\rho}+2\right)$ in the final line can be made arbitrarily small by picking ρ close to one. Moreover, the inequality $\Delta < p\frac{1+p_1}{2+p_1}$ ensures that the term $\left(1-p_1+2p_1^2-2\Delta-p_1\Delta\right)\frac{p_1c_2}{\Delta}-(1-p)^2\Delta$ is positive for c_2 sufficiently close to $\frac{\Delta^2}{p_1}$. Therefore, $\tilde{D}$ is positive for ρ and c_2 sufficiently close to their respective upper bounds.

To see renegotiation proofness, note that (3) can be simplified into

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

It is straightforward algebra to show that, as $c_1 \leq \bar{c}_1$ this hold whenever c_2 is close to $\bar{c}_2$ and ρ is close to 1.

Proof of Proposition 3 Similar to Proposition 1, we only need to consider the five effort profiles stated there. As $\frac{1}{2}\min\left\{p_1, p_1+3-\sqrt{p_1^2-6p_1+13}\right\} < p_1\frac{1+p_1}{2+p_1}$, the quota contract is renegotiation-proof and outperforms the linear contract under the conditions stated in the proposition statement. It remains to show that (a) there is no other renegotiation-proof contracts that induces the effort profile (ii) and give the principal a higher expected profit; (b) there is no other renegotiation-proof contracts that induces the effort profiles (iii) to (v) and give the principal a higher expected profit.

Consider statement (a). The optimal quota contract identified in Lemma 4 induces ONLY the agent with $y_1 = 1$ to withhold output. The other two possibilities are inducing: (1) gaming by only the agent with $y_1 = 2$; and (2) gaming by the agent with EITHER $y_1 = 1$ OR $y_1 = 2$.

For the first possibility, it is without loss to consider the following form of contract:

$$\begin{aligned}
&b_1(0) = 0, b_1(1) = B; b_1(2) = B > 0;\\
&b_2(3;0) \equiv \beta; b_2(y_2;0) = 0 \text{ for } y_2 = 0,1,2;\\
&b_2(1;1) = c_2/\Delta; b_2(0;1) = 0.
\end{aligned}$$

We show that the contract is NOT renegotiation-proof. Under this contract, the agent with $y_1 = 1$ turns in the output at the end of $t = 1$, while the agent with $y_2 = 2$ carries over two units of output to $t = 2$. First, to solicit effort in the second round of $t = 1$ following a first-round failure, requires $B \geq \frac{c_1-\rho(p_1-\Delta)c_2}{\Delta}$. Second, to induce effort in the second round of $t = 1$ following a first-round success requires $\Delta\left(\rho\left(p_1\beta-\left(\frac{p_1}{\Delta}\right)c_2\right)-B\right) \geq c_1$. The two inequalities implies $\beta \geq \frac{2c_1+\rho c_2\Delta}{p_1\rho\Delta}$. This implies that the continuation payoff of the agent with $y_1 = 2$ at the beginning of period 2 is at least

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

where the inequality follows from the assumption that $\frac{c_2}{c_1} \leq \frac{1}{\rho} \min \left\{ \frac{1}{(p_1 - \Delta)}, \frac{1}{1-(p_1-\Delta)} \right\} < \frac{2}{\rho}$. Therefore, at $t = 2$, the principal can screen the agent $y_1 = 0$ from $y_1 = 2$ without incurring any information rent.

For the second possibility, it is without loss to consider the following form of contract:

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

We show that the contract is more costly to the principal than the optimal quota contract. First, inducing effort in $t = 2$ for the agent with $y_1 = 2$ requires $B \geq \frac{c_2}{\Delta} + \beta$, whereas inducing effort in the second round of $t = 1$ following a first-round failure in $t = 1$ requires $\beta \geq \frac{c_1 + \Delta c_2}{\rho p_1 \Delta} \equiv \beta^*$ (as defined in Proposition 4). It is straightforward algebra to check that

$$B^*(\beta^*) + \delta p \frac{c_2}{\Delta} < \rho \left(\beta^* + p_1 \frac{c_2}{\Delta} \right),$$

implying that the expected agency cost is lower under the optimal quota contract than the alternative contract considered here.

Now consider statement (b). First, to induce effort profile (iii) under the renegotiation-proofness constraint, it is without loss to consider the following form of contract:

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

We show that the contract is NOT renegotiation-proof. As effort is induced at the second round following a first-round failure in $t = 1$, but not in the subsequent $t = 2$, it is necessary that $\beta \in [\frac{c_1}{\rho(p_1-\Delta)\Delta}, \frac{c_2}{\Delta})$. To induce effort in the second round of $t = 1$ following a first-round success requires $\Delta\rho((1-p_1)\beta + p_1 B - c_2) \geq 2c_1$, which implies a continuation payoff of an agent with $y_1 = 2$ is at least $\frac{2c_1}{\rho\Delta}$. However, as $\frac{2c_1}{\rho\Delta} > \frac{c_2}{\Delta}$, the principal can screen the agent with $y_1 = 2$ from the agent with $y_1 = 1$ without incurring any information rent by offering a contract with $b_2^1(2) = \frac{c_2}{\Delta}$ and $b_2^1(1) = 0$.

Second, we show that any contract that induces effort profile (iv) is less profitable than the optimal quota contract. Using the notation from the proof of Proposition 1, it suffices to show that $\Pi^T - L > \Pi^{(iv)}$. To this end, consider

$$
\begin{aligned}
&\Pi^T - \Pi^{(iv)} - L \\
&\geq \frac{p_1}{\Delta}\left[\delta\left(1-p_1\right)\Delta\left(\frac{\Delta^2}{p_1} - c_2\right) + \left(p_1 - \Delta\right)\left(1-\Delta\right)p_1 c_2\right] \\
&-\left(\rho p_1^2\left(2\Delta - \left(2p_1 - 1\right)\right)\frac{c_2}{\Delta} + \frac{1}{2}\left(1-\rho\right)\left(1 + \frac{c_1}{\rho\Delta} - \left(p_1 - \Delta\right)\frac{c_2}{\Delta}\right)\right) \\
&= \delta\left(1-p_1\right)\Delta^2 + \left(-\delta\left(1-p_1\right)\Delta + \left(p_1 - \Delta\right)\left(1-\Delta\right)p_1 - \rho p_1\left(2\Delta - \left(2p_1 - 1\right)\right)\right)\frac{p_1 c_2}{\Delta} \\
&-\frac{1}{2}\left(1-\rho\right)\left(1 + \frac{c_1}{\rho\Delta} - \left(p_1 - \Delta\right)\frac{c_2}{\Delta}\right) \\
&\geq \left[\delta\left(1-p_1\right)\Delta^2 + \left(-\delta\left(1-p_1\right)\Delta + \left(p_1 - \Delta\right)\left(1-\Delta\right)p_1 - \rho p_1\left(1 - 2\left(p_1 - \Delta\right)\right)\right)\frac{p_1 c_2}{\Delta}\right] \\
&-\frac{1}{2}\left(1-\rho\right)\left(1 + \frac{2}{\rho}\right).
\end{aligned}
$$

The first inequality makes use of results from the proofs of Proposition 1 and Proposition 2. The second inequality makes use of $c_1 \leq \bar{c}_1$ and $p_1 > \frac{1}{2}$. As $\frac{1}{2}\left(1-\rho\right)\left(1 + \frac{2}{\rho}\right)$ can be made arbitrarily small by choosing ρ close to 1, it suffices to show that the term in bracket in the final line is strictly positive for c_2 close to $\frac{\Delta^2}{p_1}$ and ρ close to 1. As the term is linear in c_2 and positive at $c_2 = 0$, it suffices to show that it is strictly positive at $c_2 = \bar{c}_2$ and $\rho = 1$, or equivalently, $\Delta^2 - \left(3 + p_1\right)\Delta + \left(3p_1 - 1\right) > 0$. On solving, $\Delta < \frac{1}{2}\left(p_1 + 3 - \sqrt{p_1^2 - 6p_1 + 13}\right)$.

Finally, we show that any contract that induces effort profile (v) is less profitable than the optimal quota contract. Using the notation from the proof of Proposition 1, it suffices to show that $\Pi^T - L > \Pi^{(v)}$. To this end, consider

$$
\begin{aligned}
&\Pi^T - \Pi^{(v)} - L \\
&\geq 2\Delta\left(1-p_1\right)\rho\left(p_1 - \Delta\right) - \left(\rho p_1^2\left(2\Delta - \left(2p_1 - 1\right)\right)\frac{c_2}{\Delta} + \frac{1}{2}\left(1-\rho\right)\left(1 + \frac{c_1}{\rho\Delta} - \left(p_1 - \Delta\right)\frac{c_2}{\Delta}\right)\right) \\
&\geq \rho\left[2\Delta\left(1-p_1\right)\left(p_1 - \Delta\right) - p_1^2\left(1 - 2\left(p_1 - \Delta\right)\right)\frac{c_2}{\Delta}\right] - \frac{1}{2}\left(1-\rho\right)\left(1 + \frac{2}{\rho}\right).
\end{aligned}
$$

The first inequality makes use of results from the proofs of Proposition 1 and Proposition 2. The second inequality makes use of $c_1 \leq \bar{c}_1$ and $p_1 > \frac{1}{2}$. As $\frac{1}{2}\left(1-\rho\right)\left(1 + \frac{2}{\rho}\right)$ can be made arbitrarily small by choosing ρ close to 1, it suffices to show that the term in bracket in the final line is strictly positive for c_2 close to $\frac{\Delta^2}{p_1}$ and ρ close to 1, or equivalently, $\Delta < \frac{p_1}{2}$. Q.E.D.