Incentives for Quality through Endogenous Routing

Lauren Xiaoyuan Lu · Jan A. Van Mieghem · R. Canan Savaskan

Kellogg School of Management, Northwestern University

July 14, 2006. Revised January 5, 2007

Abstract

We study how rework routing together with wage and piece rate compensation can strengthen incentives for quality. Traditionally, rework is assigned back to the agent who generates the defect (in a *self routing* scheme) or to another agent dedicated to rework (in a *dedicated routing* scheme). In contrast, a novel *cross routing* scheme allocates rework to a parallel agent performing both new jobs and rework. The agent who passes quality inspection or completes rework receives the piece rate paid per job. We compare the incentives of these rework allocation schemes in a principal-agent model with embedded quality control and routing in a multi-class queueing network. We show that conventional self routing of rework can never induce first-best effort. Dedicated routing and cross routing, however, strengthen incentives for quality by imposing an implicit punishment for quality failure. In addition, cross routing leads to workload allocation externalities and a prisoner's dilemma, thereby creating highest incentives for quality. Firm profitability depends on capacity levels, revenues, and quality costs. With ample capacity, dedicated routing and cross routing both achieve first-best profit rate, while self routing does not. With limited capacity, cross routing generates the highest profit rate when appraisal, internal failure, or external failure costs are high, while self routing performs best when gross margins are high. When the number of agents increases, the incentive power of cross routing reduces monotonically and approaches that of dedicated routing.

Key words: queueing networks; routing; Nash equilibrium; quality control; piece rate; epsilon equilibrium.

1 Introduction

This paper investigates how rework routing together with wage and piece rate compensation can strengthen incentives for quality and improve firm profits in a setting where agents "compete" for rework. It is motivated by the practice of the service operations firm Memphis Auto Auction, which is a wholesale automotive liquidator of used vehicles that employs two teams of employees

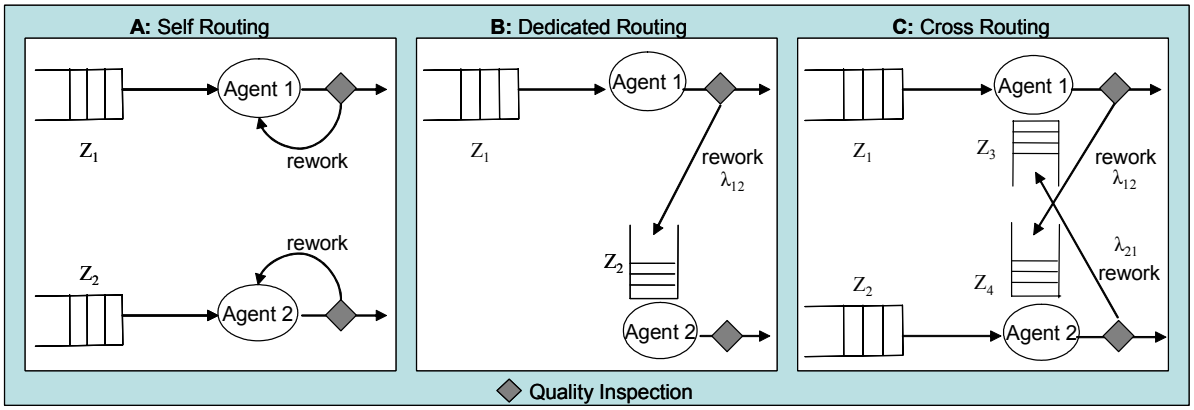

Figure 1: Three Rework Routing Schemes: (A) Self Routing, (B) Dedicated Routing, and (C) Cross Routing.

to clean and detail vehicles in parallel. The employees are paid piece rates only if their jobs pass quality inspection while the quality control leader is paid salary plus a bonus based on overall work quality. The firm ties compensation to quality through an unconventional rework routing scheme illustrated in Figure 1C that we call *cross routing*. This cross routing of rework contrasts with the two traditional practices that assign rework back to the team who generates the defect (Figure 1A) or to a dedicated rework team (Figure 1B) and also pay piece rate only to the team whose job passes quality inspection. We shall show that these three rework routing policies generate different first-pass quality incentives and that the "competition" for rework implicit in cross routing can yield superior outcomes.

Our main research goal is to explain how these three routing and incentive schemes compare in terms of quality and firm profits. Our analysis uses a principal-agent model with endogenous piece rate, quality control, and routing in a multi-class queueing network. Rework routing impacts agent incentives to exert quality-improving effort in two important ways. First, self routing gives agents a second chance to work on a job and earn the piece rate, resulting in a disincentive to exert first-pass effort. In contrast, dedicated routing and cross routing implicitly punish the agents for quality failure by allocating rework (and thus the associated piece rate) to another agent, thereby boosting the incentives for first-pass quality.

Second, whereas self routing gives each agent independent and direct control over the workload of new jobs and rework, the workload in cross routing is determined by the equilibrium outcome of the noncooperative effort game played between the two agents. When rework takes less effort than new jobs, rework is preferred, which prompts the agents to increase their first-pass effort as a result

of the *workload allocation externality* arising from the effort game. To illustrate this externality, consider the strategic interaction between the agents and the flow dynamics of the queueing network. In a capacity constrained system, both agents are continuously busy working on either new jobs or rework. To receive more rework, agent 1 increases first-pass effort and sends less rework to agent 2. Keeping his effort unchanged, agent 2 then automatically processes more new jobs and sends more rework to agent 1. Suffering from reduced pay due to low rework inflow, however, agent 2 increases first-pass effort to counteract. Consequently both agents exert high effort and receive low rework allocation in equilibrium. This equilibrium exhibits a *prisoner's dilemma*, where each agent has an incentive to exert high effort when the other agent exerts low effort, even though the agents would jointly benefit from a cooperative outcome of both exerting low effort (i.e., the effort level under self routing). This shows why noncooperative behavior is crucial for cross routing: when the agents collude, it is equivalent to self routing. We loosely refer to this noncooperative behavior as competition.

While higher first-pass effort produces fewer internal and external defects, it does not always lead to higher profits for the principal. On one hand, inducing first-pass effort benefits the principal by improving quality and reducing three of the four quality costs in Juran's cost-of-quality framework (Juran & Gryna (1993)): internal failure costs, external failure costs, and appraisal costs. On the other hand, excessively high first-pass effort lowers throughput[1] as the agents spend more effort (processing time) per job on average. Since piece rate compensation cost can be deemed as a form of prevention costs, our model covers all of the four dimensions of the cost-of-quality framework. It predicts that the principal would strive for the optimal defect rate (which has a one-to-one relationship with the induced first-pass effort) to achieve the lowest costs by balancing the cost of non-conformance with appraisal and prevention costs. Built on this cost minimization view of quality management, our model adds an additional dimension: throughput and thus revenues also impact a firm's quality control policies.

In our model, effort is the first-pass service time chosen by the agents. Both quality and throughput are determined by the effort choices. Compared to the celebrated multitask principal-agent model of Holmstrom & Milgrom (1991), this model embeds the intrinsic trade-off between quantity and quality in a single dimensional decision variable. This is appropriate when the quantity effort and quality effort are not separable, as displayed in Figure 2.

When service time is observable and also contractible, the moral hazard problem is eliminated.

[1] The total expected service time (including rework time) is convex in first-pass effort.

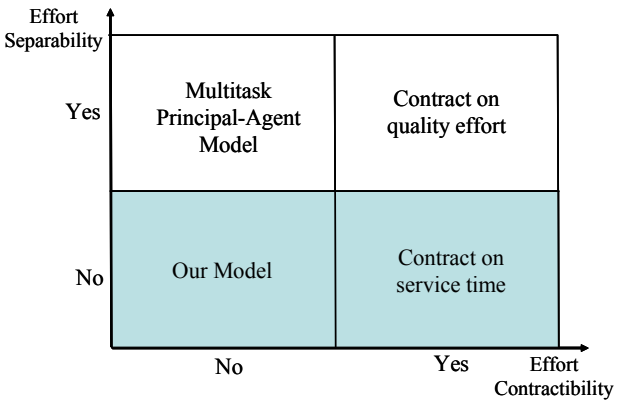

Figure 2: The effort contractibility and separability matrix.

Indeed, in a long-run repeated setting, the principal may fully infer the average service time from the throughput of the agent taking into account of quality loss. Though service time may be observable, directly contracting on it can be problematic. For example, in a call center, it is uncommon to tie agents' compensation to their handle time per call. The difficulty for contracting on handle time is two fold: behavioral issues and agent heterogeneity. First, handle time per call is highly variable so that contracting on them would require a contract involving probability distribution. However, statistical thinking requires mental sophistication and effort, making a probabilistic contract impractical to implement. If the call center persists on contracting on a target handle time, agents may be misguided in terms of the "ideal" time to spend on individual calls. Because of the challenge of statistical thinking, agents may not be able to treat the target as an average and would try to meet the target time for every single call. According to Gans, Koole & Mandelbaum (2003), an incentive scheme that rewards agents for maintaining low average handle time can lead them to hang up on customers. Second, agents are heterogenous in experience and skill levels, adding to the challenge of specifying an appropriate target handle time. Therefore, contracting on handle time is precluded for an operationally excellent call center.

Under dedicated routing and cross routing, assigning rework to a different agent implicitly punishes the agent for shirking. Such punishment could be replicated by a modified self routing scheme where the principal executes a monetary punishment whenever a defect is identified. Similarly, she could also implement a bonus that is paid whenever a job passes quality inspection in the first pass. Both contracts can achieve first best effort and profits. However, these contracts are not the focus of this paper. In real life, it is difficult for a principal to "force" rework without or with negative pay.

In contrast, cross routing of rework is a more "fair" contract in the sense that the principal always pays the full piece rate per job, but only rewards the agent whose work passes quality inspection. In a call center, cross routing has the additional benefit of higher utilization of agents because unresolved issues can be routed to any available agent rather than to the first-contact agent.[2]

From a methodology perspective, restricting the principal's contract space to wage plus piece rate enables us to isolate the incentive effects brought by the different rework routing schemes. The economic literatures on agency and contract theory mostly focus on incentives using purely monetary instruments. We take those economic schemes as our benchmark and investigate whether adding operational instruments (e.g., rework routing in our case) can achieve similar or even superior outcomes. As such our work marries economics with operations strategy and processing network design.

Given that quality output and inspection are imperfect, endogenous probabilistic routing of rework is a central feature of our model. Because the agents' actual workload allocation of new jobs and rework is endogenously determined by their effort, a natural tool to characterize the effort equilibrium arising in the flow system is queueing theory. A deterministic model cannot capture the inherent variability and flow dynamics of the system. The endogeneity of rework routing also has a second aspect: The principal compares the financial performance of the three routing schemes and chooses the most profitable one.

Using an analytical model we establish the following results. Traditional self routing of rework can never induce agents to exert first-best effort. Dedicated routing and cross routing, however, offer some remedy by inducing higher effort and quality, which can lead to higher profits for the principal. As a result, piece rates paid in these two schemes are generally higher when holding the wage rate constant. Firm profitability depends on capacity levels, revenues, and quality costs. With ample capacity, dedicated routing and cross routing both achieve first-best profit rate, while self routing does not. With limited capacity, cross routing generates the highest profit rate when appraisal, internal failure, or external failure costs are high, while self routing performs best when gross margins are high. When the number of agents increases, the incentive power of cross routing reduces monotonically under certain conditions and approaches that of dedicated routing.

The remainder of the paper is organized as follows. Section 2 reviews related literature while

[2] We learned that at Dell Inc.'s call center for corporate IT service, a credit is awarded to agents for each customer call that does not fail in seven days after service. Otherwise, it is routed to any agent available, who is eligible to earn the credit.

Section 3 lays out the main model. Sections 4 and 5 analyze the networks with ample capacity and limited capacity, respectively. In each of the two sections, we first derive the first-best benchmark and then analyze the three rework routing schemes, and finally compare their performance. Section 6 analyzes a project management setting while Section 7 concludes. In the rest of the paper, we will use superscripts FB, S, D, and C to denote solutions for first best, self routing, dedicated routing, and cross routing, respectively.

2 Related Literature

This paper contributes to three streams of literature. The first stream is the economics literature on compensation and job design, which studies the moral hazard problem that arises when an agent's effort is imperfectly observed. Compensation is thus often based on output instead of effort. Holmstrom & Milgrom (1991) explain the trade-offs between inducing effort towards quantity vs. quality with a multitask principal-agent model. In their model, producing high volume and good quality is viewed as two tasks of an agent's job. They argue that it would be costly, if not impossible, to achieve good quality with piece rate compensation if quality were poorly measured. Instead of taking a multitask approach, we manifest the intrinsic trade-off between quantity and quality by a single dimensional decision variable, i.e., the average processing time spent per job. Moreover, we provide theoretical support that smart routing of rework is capable of inducing quality-improving effort even under piece rate compensation. Lazear (2000) provides empirical evidence that piece rate compensation significantly improves productivity. In Lazear's real-world example, rework is assigned to the originating agent (i.e., self routing) and quality does not deteriorate after the firm implements piece rate compensation. He argues that the employees have incentives to get it right the first time because rework is costly. In contrast, we will show that agents always exert system suboptimal quality effort under self routing.

Holmstrom & Milgrom (1991) also demonstrate that job design is an important instrument for the control of incentives. They find that tasks should be grouped such that easily measured tasks are assigned to one agent and hard-to-measure tasks to the other. Though we use a one-dimensional principal-agent model, there are two tasks in our model that differ in their measurability: first-pass work is monitored imperfectly by quality inspection while rework is assumed to have no uncertainty in quality. Supporting Holmstrom & Milgrom (1991)'s theory that tasks should be separated according to their measurability characteristics, we show that dedicated routing achieves

advantageous incentive power over self routing.

The second relevant stream of literature is on the economics of quality control and inspection in a game-theoretic setting. These papers mostly consider quality-related contractual issues between firms and are only tangentially related to our work. For example, Reynier & Tapiero (1995) study the effect of contract parameters and warranty costs on the choice of quality by a supplier and the quality control policy by a buyer. Baiman, Fischer & Rajan (2000) focus on how contractibility of quality-related information impact the product quality and profits of a supplier and a buyer. Our work studies how rework routing and costs of quality affect the employees' choice of quality-improving effort and a firm's quality inspection policy.

From a methodological perspective, we combine the two previous literatures on principal-agent models and quality management with that of network flows in general, and queueing networks in particular. Much of agency theory seeks contracts that maximize a principal's objective subject to an agent's post-contractual opportunistic behavior. However, little is known about quality control policies, i.e., how precisely should performance be measured? Queueing network models can capture system dynamics and quality inspection levels and allow us to draw operational insights that are largely missing in the existing agency literature. By considering capacity-constrained systems, we allow agents' effort levels (i.e., processing times) to directly impact system throughput, i.e., productivity. Similar work can be found in the literature that studies the impact of decentralized decision making on process performance in queueing systems. Seminal work by Naor (1969) studies how pricing can achieve social optimum and prevent performance degradation as a result of customers' self-interested behavior. Many followers (e.g., Mendelson & Whang (1990), Van Mieghem (2000), Ha (2001), etc.) also design pricing mechanisms to achieve system optimal performance, but none of these works model quality inspection and rework.

Principal-agent models in queueing systems have been explored in the operations management literature. A sample of recent papers include Gilbert & Weng (1998), Plambeck & Zenios (2000), Shumsky & Pinker (2003), Gunes & Aksin (2004), Benjaafar, Elahi & Donohue (2006), Cachon & Zhang (2006), and Ren & Zhou (2006). Plambeck & Zenios (2000) study incentives in a dynamic setting where an agent's effort influences the transition probabilities of a system. Similarly in our model, probabilistic routing is determined by agents' effort. But, our model captures system dynamics resulting from the strategic interaction between agents, which is not present in Plambeck & Zenios (2000). Our paper is closely related to Shumsky & Pinker (2003) in that the principal designs incentives to induce effort in steady state, but differs in two important ways: First, we

explicitly model the queueing network dynamics and also consider the case where the system is capacity constrained. Second, the principal in our model hires two agents whose expected utility rates are interdependent. Therefore, we need to investigate agents' strategic interactions and derive the effort Nash equilibrium. Gunes & Aksin (2004) model the interaction of market segmentation, incentives, and process performance of a service-delivery system using a single-server queue embedded in a principal-agent framework. The novelty of our model lies in that we model two endogenous queues, i.e., the rework queues that are generated by the agents and the arrival rate of rework is endogenously determined by the agents' effort. Ren & Zhou (2006) use a multitask principal-agent model to study the coordination of capacity and service quality decisions in call center outsourcing.

Sharing a common theme with Gilbert & Weng (1998), Cachon & Zhang (2006), and Benjaafar et al. (2006), our paper uses an operational instrument to create incentives in a multi-agent processing system. In all these papers service rate choice may be observable to the principal but is not directly contracted upon. Rather, demand allocation schemes (in our case, rework routing allocation schemes) are used to create competition (in our case, to induce workload allocation externalities and thus create incentives). Multi-agent games in queueing systems can also be found in Cachon & Harker (2002) and Parlakturk & Kumar (2004), whose models, however, do not involve a principal.

In our motivating example, Memphis Auto Auction employs teams to complete jobs. In this paper, we will treat teams as agents and ignore the intra team incentive issues that may arise due to free riding and collaboration. A relevant reference for team incentives is Hamilton, Nickerson & Owan (2003), which empirically investigates the impact of teams on productivity. They distinguish individual piece rate used in flow production from group piece rate used in modular production. They find that group piece rate has a stronger incentive effect on productivity than individual piece rate due to collaboration among team members.

3 The Model

Operational Flows. Consider an operation where a principal hires two identical risk neutral agents to complete work ("jobs") and subsequently inspects their output quality. The principal sets quality inspection precision $p \in [0, 1]$, which is the probability of catching a defect given a bad output. (A good quality output passes inspection with probability 1.) This inspection precision is observable to the agents. p can be interpreted as a sampling frequency of inspection. We assume

that the principal commits to p once announced.

Each agent chooses first-pass effort (service time) t, where $t \geq \underline{t}$ and $\underline{t} > 0$ is the minimum effort that can be exerted. The minimum effort assumption prevents the extreme case where the agents directly move jobs to quality inspection and always do rework. We assume that the agents' service time of each job has mean t. This strategic decision variable drives the output quality. We assume that the agents adopt open-loop strategies and thus the service time decision is one shot.

Let $F(t)$ denote the probability of producing good quality given first-pass effort t, with $F(\underline{t}) = 0$ and $F(\infty) = 1$. We assume that F is strictly concave and increasing (i.e., $F'' < 0$, $F' > 0$), reflecting decreasing return on effort, and denote $f = F'$ and $\bar{F} = 1 - F$. Upon identifying defects, the principal routes the rework either to the originating agent in self routing, to the agent dedicated to rework in dedicated routing, or to the parallel agent in cross routing. We assume that rework always generates good output, thus poor quality only results from not catching the first-pass defects. The overall quality conformance level that an external customer experiences is $Q = F(t) + p\bar{F}(t)$.

We will show that the incentive effects of the three routing schemes crucially depend on whether the network is capacitated. With ample capacity, each agent is supplied with a renewal process of job arrivals. In steady state, the agents have idle time and the throughput is driven by exogenous demand arrival rate. In contrast, when the system has limited capacity, the agents are continuously busy and their effort levels directly impact throughput. The motivation for looking at both cases is that they each represent a different real-world operating system. A system with limited capacity models a make-to-stock system where agents are scheduled to complete jobs that are either planned or have arrived previously in bulks. In such systems, the agents are fully utilized (for a fixed period of time) and response time is often not a critical issue. In contrast, a system with ample capacity models a make-to-order system where response time is critical.

For tractability, we assume that rework takes r units of time on average, where r is common knowledge. Since defects have to be corrected as instructed by the principal, we assume that rework effort is contractible, i.e., no moral hazard problem in rework. We argue that even if agents may exhibit opportunistic behavior in performing rework, the effect is limited because identified defects have to be corrected completely. Furthermore, rework has preemptive priority over new jobs. This priority rule is adopted because of two considerations. First, in a capacitated system, agents can be always engaged in new jobs. Without the priority rule, defects may never be reworked. Second, the priority rule simplifies analytics of the model. Finally, we assume that rework takes less time

than the minimum first-pass effort:

$$r \leq \underline{t}. \tag{A1}$$

This assumption allows us to focus on the interesting range of parameter values that highlight the moral hazard problem and the efficacy of "smart" rework routing in inducing effort. We will discuss the implications of this assumption when comparing the performance of the three routing schemes in Section 4.3.

We assume that service times and inter-arrival times are independent and exponentially distributed. Though not considered in this paper, this assumption allows for response time evaluation. We shall note that these assumptions can be relaxed and that a renewal process for both job arrival and service times is sufficient for all subsequent results to hold.[3]

Financial Flows and Incentives. Each agent earns wage rate w, and in addition piece rate b when completing a new job that passes quality inspection or when completing a rework. Under dedicated routing, the two agents earn different wage rates w_1 and w_2. Their average is the equivalent wage rate for comparison purpose. The agents' disutility of effort per unit time is a. Without loss of generality, we normalize the agents' reservation utility to be 0. In a competitive labor market, a can also be interpreted as the outside wage rate. The principal earns gross margin v per completed job that passes quality inspection, pays agents, and incurs three quality costs classified as in Juran's cost-of-quality framework: (1) appraisal cost per new job $C(p)$. We assume $C(0) = C'(0) = 0$ and $C'(1) = \infty$, which implies that in equilibrium the principal chooses an interior inspection policy, i.e. $p \in (0, 1)$. In addition, $C' > 0$, $C'' > 0$. Note that these are assumptions often used in the quality management literature (e.g., Baiman et al. (2000)). (2) internal failure cost per new job c_I. (3) external failure cost per new job c_E. (External failure costs are typically larger than internal failure costs: $c_E > c_I$. Otherwise, the principal would have no incentives to fix defects internally.) We assume that the principal maximizes her long-run average profit rate per agent, denoted by V, while the agents maximize their long-run average utility rate, denoted by U.

4 Incentives and Routing with Ample Capacity

Ample capacity implies that the principal maintains a sufficient staffing level to complete all jobs with appropriate response time, and that the agents have idle time in steady state. Hence, the throughput of the system is driven by the exogenous market demand, which is represented by the

[3] We thank Harry Groenevelt for pointing this out.

mean arrival rate of jobs per agent and denoted by λ. The principal focuses on reducing internal and external failure costs through quality inspection and inducing first-pass effort while controlling for appraisal costs and compensation costs. Let ρ_i denote the utilization of agent i. Throughout this section, we assume that the system is stable in steady state. The stability condition is $\max_{i\in\{1,2\}} \rho_i < 1$.

4.1 The First-Best Benchmark (Contractible Effort)

When effort t is contractible, the principal's problem is independent of whether rework is performed by the originating agent or a different agent. For expositional convenience, we derive the first-best benchmark using the self routing scheme. The agents spend on average $t + p\bar{F}(t)r$ time units per job. Since the job arrival rate is λ per agent, renewal theory yields that the agents' long-run average utility rate is $\lambda[b-a(t+p\bar{F}(t)r)]-w$. Though the principal hires two agents, the contracting problem of each agent is independent and identical. The principal maximizes

$$\begin{aligned} V^{FB} &= \max_{0\le p\le 1, t\ge \underline{t}, w, b} \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E)) - C(p)] - w, \\ \text{subject to} &\quad \lambda[b - a(t + p\bar{F}(t)r)] + w \ge 0 \quad \text{(IR)}. \end{aligned}$$

The individual rationality (IR) constraint specifies the agents' outside option. Note that the IR constraint is identical for both agents. Since the principal's profit rate is monotone decreasing in w and b, the IR constraint must bind, simplifying the principal's problem to an optimization problem of two variables: t and p. Let $\{t^{FB}, p^{FB}\}$ denote the first-best solution.[4] Since $\rho_i = \lambda(t^{FB} + p\bar{F}(t^{FB})r)$, the stability condition becomes $\lambda < \frac{1}{t^{FB}+p\bar{F}(t^{FB})r}$.[5] For a stable system, Lemma 1 characterizes the first-best solution (all proofs are relegated to the Appendix).

Lemma 1 *If $c_E > c_I + ar > \frac{a}{f(\underline{t})}$, there exists an interior first-best solution $\{t^{FB}, p^{FB}\}$ given by*

$$f(t^{FB}) = \frac{1}{p^{FB}r + \frac{1}{a}(p^{FB}c_I + (1-p^{FB})c_E)}, \tag{1}$$

$$C'(p^{FB}) = \bar{F}(t^{FB})(c_E - c_I - ar). \tag{2}$$

The optimal effort depends on the density of the probability function of producing good quality. This is because the agents face an increasing concave "production function" $F(\cdot)$, the density of which measures the marginal return of effort. It is simple to show that $\frac{\partial^2 V}{\partial t \partial p} < 0$, i.e., t and p are

[4] We ignore the issue of uniqueness of solution as all of our subsequent results hold for any interior optimum.

[5] Obviously, t^{FB} is a function of λ and F. Only for specific instances of F can this inequality be solved explicitly in terms of model primitives.

strategic substitutes. The principal can select from infinite pairs of wage rate and piece rate to satisfy the IR constraint at equality. Since the principal is the Stackelberg leader and the agent earns zero utility rate in equilibrium, the principal's objective is identical to a central planner's. Therefore, the first-best solution achieves the Pareto optimum for the entire system.

4.2 Optimal Incentives for the Three Networks (Noncontractible Effort)

4.2.1 Self Routing

When effort t is not contractible and the rework is routed back to the originating agent, the principal maximizes

$$\begin{aligned} V^S &= \max_{0\le p\le 1, w, b} \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E)) - C(p)] - w, \\ \text{subject to} \quad & \lambda[b - a(t + p\bar{F}(t)r)] + w \ge 0 \quad \text{(IR)}, \\ & t \in \arg\max_{t' \ge \underline{t}} \lambda[b - a(t' + p\bar{F}(t')r)] + w \quad \text{(IC)}. \end{aligned}$$

The additional incentive compatibility (IC) constraint describes the agents' post-contractual optimization behavior. Since the two agents are completely independent and symmetric, we only need a single IR and IC constraint.

Lemma 2 *With ample capacity and self routing, the agents' unique optimal effort is*

$$t^S = \begin{cases} f^{-1}(\frac{1}{pr}) & \text{if } f(\underline{t}) > \frac{1}{pr} \\ \underline{t} & \text{if } f(\underline{t}) \le \frac{1}{pr} \end{cases}. \tag{3}$$

Since the agents have sufficient time to complete all jobs and always earn the piece rate of each job, the agents' optimal effort is not impacted by the job arrival rate λ and the piece rate b. However, the first-pass effort increases when the principal raises the quality inspection precision or when rework is costly to the agents. The stability condition becomes $\lambda < \frac{1}{t^S + p\bar{F}(t^S)r}$.

4.2.2 Dedicated Routing

Without loss of generality, we assign new jobs to agent 1 and rework to agent 2. To keep the system's supply of jobs unchanged, agent 1 is assigned with job arrival rate 2λ. The principal

maximizes

$$\begin{aligned} V^D &= \max_{0\le p\le 1, w, b} \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E)) - C(p)] - \frac{w_1 + w_2}{2}, \\ \text{subject to} \quad & 2\lambda[(1 - p\bar{F}(t))b - at] + w_1 \ge 0 \quad \text{(IR1)}, \quad 2\lambda p\bar{F}(t)(b - ar) + w_2 \ge 0 \quad \text{(IR2)} \\ & t \in \arg\max_{t' \ge \underline{t}} 2\lambda[(1 - p\bar{F}(t'))b - at'] + w_1 \quad \text{(IC1)}. \end{aligned}$$

Since agent 2's rework effort is contractible, only IC1 is needed.

Lemma 3 *With ample capacity and dedicated routing, agent 1's unique optimal effort is*

$$t^D = \begin{cases} f^{-1}(\frac{a}{pb}) & \text{if } f(\underline{t}) > \frac{a}{pb} \\ \underline{t} & \text{if } f(\underline{t}) \le \frac{a}{pb} \end{cases}.$$

Now agent 1's optimal effort depends on both p and b. Therefore, the principal can induce higher first-pass effort not only by increasing the quality inspection precision but also by raising the piece rate. Agent 1 and 2's utilizations are $\rho_1 = 2\lambda t^D$ and $\rho_2 = 2\lambda p\bar{F}(t^D)r$, respectively. The stability condition becomes $\lambda < \min\{\frac{1}{2t^D}, \frac{1}{2p\bar{F}(t^D)r}\} = \frac{1}{2t^D}$.

4.2.3 Cross Routing

When rework is assigned to the parallel agent, a rework queue is generated and its queue size depends on the first-pass effort of the originating agent. We now must characterize the rework equilibrium queues as part of the principal-agent incentive problem. For the multi-class queueing network illustrated in Figure 1C, we define the following rates for $i, j \in \{1, 2\}$ and $i \ne j$:

- Agent $i's$ new job service rate $\mu_i^n = \frac{1}{t_i}$
- Agent $i's$ defect generation rate (or agent $j's$ rework arrival rate) $\lambda_{ij} = \frac{p\bar{F}(t_i)}{t_i}$
- Rework service rate $\mu^r = \frac{1}{r}$ (same for both agents)

Let a four-dimensional vector (Z_1, Z_2, Z_3, Z_4) represent the state of the four queues of the system (two new job queues and two rework queues). The detailed balance equations are too complex to be solved analytically in closed form. However, we do not need the limiting distribution of every single state to compute the utility rate of the agents. It suffices to know the aggregate probabilities of the agents being idle π_i^0, working on new jobs π_i^n, and working on rework π_i^r. In steady state,

the queueing network must satisfy

$$\begin{aligned} \pi_i^0 + \pi_i^n + \pi_i^r = 1 \quad & \text{(Law of total probability)}, \\ \lambda = \mu_i^n \pi_i^n \quad & \text{(Balance of agent } i's \text{ new job queue)}, \\ \lambda_{ji} \pi_j^n = \mu^r \pi_i^r \quad & \text{(Balance of agent } i's \text{ rework queue)}, \end{aligned}$$

for $i, j \in \{1, 2\}$ and $i \neq j$. Solving the above equations yields

$$\pi_i^0 = 1 - \lambda t_i - \lambda p \bar{F}(t_j) r, \quad \pi_i^n = \lambda t_i, \quad \pi_i^r = \lambda p \bar{F}(t_j) r$$

Thus, agent $i's$ long-run average utility rate

$$\begin{aligned} U_i(t_i, t_j) &= \pi_i^n \times \frac{(1 - p\bar{F}(t_i))b - at_i}{t_i} + \pi_i^r \times \frac{b - ar}{r} + w \\ &= \lambda[(1 - p\bar{F}(t_i))b - at_i] + \lambda p \bar{F}(t_j)(b - ar) + w. \end{aligned}$$

Notice that the first term is agent i's average reward rate from working on new jobs while the second term is his average reward rate from completing rework generated by agent j.

Lemma 4 *With ample capacity and cross routing, the unique Nash equilibrium of the agents' effort game is* (t^C, t^C)*, where* $t^C = t^D$ *as in Lemma 3.*

Surprisingly, the agents' optimal effort in equilibrium is independent of each other's effort and is solely determined by the principal's quality inspection and incentive decisions. Because the agents have idle time in steady state, performing rework simply reduces idle time, but does not impact their workload of new jobs. Therefore, cross routing imposes no additional effect on the incentives other than taking away the second opportunity to work on a job. This effect is also present in dedicated routing, rendering identical optimal effort in both schemes. Moreover, the two agents have no strategic interactions and behave symmetrically. Since agent i's utilization $\rho_i = \lambda(t_i + p\bar{F}(t_j)r)$, the stability condition becomes $\lambda < \frac{1}{t^C + p\bar{F}(t^C)r}$. The principal maximizes

$$\begin{aligned} V^C &= \max_{0 \leq p \leq 1, w, b} \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E)) - C(p)] - w, \\ \text{subject to} \quad & \lambda[(1 - p\bar{F}(t))b - at] + w \geq 0 \quad \text{(IR)}, \\ & t = t^C \quad \text{(IC)}. \end{aligned}$$

4.3 Comparing the Three Networks: Implicit Punishment

Comparing Equation (1) with (3) allows us to illustrate the importance of assumption (A1). Notice that when r is large, the difference between $f(t^{FB})$ and $f(t^S)$ becomes small and thus even self

routing performs close to first best. This supports the intuition that agents have incentives to get it right the first time when rework is costly. Therefore, assumption (A1) allows us to restrict our attention to the range of parameter values where agents' opportunistic behavior is prominent. In addition, small rework time makes self routing and cross routing implementable in a flow system. Otherwise, the production line has to be stopped to allow agents to complete rework backlog. The opposite extreme of the assumption is that r is sufficiently large such that $ar > c_E - c_I$. Then, it is optimal for the principal to eliminate quality inspection (or to scrap the defects) because the reduction in external failure costs cannot compensate for the high costs of internal repair.

Proposition 1 *Self routing can never implement first best. In contrast, dedicated routing implements first best with contract* $\{p^{FB}, w_1^*, w_2^*, b^*\}$ *and cross routing implements first best with contract* $\{p^{FB}, w^*, b^*\}$, *where* $b^* = ar + c_I + \frac{1-p^{FB}}{p^{FB}}c_E$, $w_1^* = 2\lambda[at^{FB} - (1 - p^{FB}\bar{F}(t^{FB}))b^*]$, $w_2^* = 2\lambda p^{FB}\bar{F}(t^{FB})(ar - b^*)$, *and* $w^* = \lambda[a(t^{FB} + p^{FB}\bar{F}(t^{FB})r) - b^*]$. *Therefore,* $V^{FB} = V^D = V^C > V^S$.

Proposition 1 reflects the weakness of the conventional self routing scheme: Because the agent has a second chance to work on a job and earn the piece rate, he has a disincentive to exert first-pass effort and takes his chance at quality inspection. This gaming behavior leads to a lower first-pass quality level, incurring higher internal and external failure costs to the principal. In contrast, both dedicated routing and cross routing can attain first best. The optimal piece rate b^* is chosen to induce the first-best effort while the wage rates are chosen to meet the agents' reservation utility rate. Further notice that $w^* = \frac{w_1^* + w_2^*}{2}$ i.e., the wage rate paid under cross routing equals the average wage rate paid under dedicated routing.

From a central planner's point of view, dedicated routing and cross routing are superior because the effort and quality inspection are set at the system optimal level. Because the agents earn their reservation utility rate under the first-best contracts, the principal achieves the highest possible profit rate under these two routing schemes. We further compare the three schemes along first-pass effort, quality output, and piece rate. We will use $y(p)$ to denote a function of p.

Corollary 1 *For all* $p \in (0, 1)$, $t^{FB}(p) = t^D(p) = t^C(p) > t^S(p)$. *Therefore,* $Q^{FB}(p) = Q^D(p) = Q^S(p) > Q^S(p)$.

Dedicated routing and cross routing provide stronger incentives for quality because assigning rework to a different agent imposes an implicit punishment on the agents for their quality failure. This punishment is derived from that the agents lose the effort spent on the jobs that fail quality

	Ample Capacity			Limited Capacity		
	effort t	piece rate b	principal's profit rate V	effort t	piece rate b	principal's profit rate V
Self	Low	Low	Low	Low	Low	Depends on
Dedicated	High	High	High	Medium	Medium	quality costs and
Cross	High	High	High	High	High	gross margins

Table 1: Comparing first-pass effort, piece rates, and profit rates of the three schemes.

inspection. Since the piece rate is paid on top of the wage rate, it can be viewed more generally as a bonus system:

$$\begin{aligned}
\text{Bonus}^S &= \text{piece rate} \times \text{inflow}, \\
\text{Bonus}^D &= \text{piece rate} \times (\text{inflow} - \text{rework sent to others}), \\
\text{Bonus}^C &= \text{piece rate} \times (\text{inflow} + \text{rework received from others} - \text{rework sent to others}).
\end{aligned}$$

The implicit punishment is monetary and is evident from the deduction of bonus under dedicated routing and cross routing.

Corollary 2 *For all $p \in (0,1)$, $b^{FB}(p) = b^D(p) = b^C(p) > b^S(p)$ at equal wage rates.*

Interestingly, we find that the piece rates paid in cross routing and dedicated routing are higher than the one paid in self routing because in the former two schemes the agents exert higher effort in equilibrium and cannot recoup the cost of effort spent on the jobs that have failed inspection. A summary of the comparison is in Table 1.

4.4 Equilibria with Many Agents

So far we have only considered an operation with two agents. It is interesting to see whether the results hold in a large operation with many agents. Let $N \in \mathbb{N}$ denote the number of agents. With self routing, the system can be scaled up proportionally because the agents are independent of each other. With dedicated routing and cross routing, however, we need to make further assumptions on how the system is scaled up and what the routing rule is. For dedicated routing, we treat each pair of agents as the basic unit and an N-agent system has $N/2$ such units (restrict N to even numbers). Obviously, the N-agent system is identical to the two-agent system in terms of profit

rate. For cross routing, we treat the agents anonymous and route rework to each one with equal probability. The long run utility rate of agent i is

$$U_i(t_i, t_{-i}) = \lambda[(1 - p\bar{F}(t_i))b - at_i] + \lambda(b - ar)\frac{p}{N-1}\sum_{j\neq i}\bar{F}(t_j) + w$$

It follows immediately that $t_i^C = f^{-1}(\frac{a}{pb})$, which is identical to that of the two-agent case. In summary, when the system has ample capacity, the system can be scaled up proportionally and the incentive effects of the three routing schemes remain unchanged.

5 Incentives and Routing with Limited Capacity

In contrast to the ample capacity case, the throughput of a capacitated system is endogenous and depends on the agents' effort. Therefore, both the agents and the principal face a trade-off between throughput and quality. Optimizing their utility rate, the agents balance the time allocated to new jobs versus rework to trade-off earning the piece rate from first-pass success with that from rework. The principal balances inducing quality-improving effort with increasing throughput. Throughout this section, we assume that the arrival rate of new jobs are sufficiently large such that the stability conditions of the previous section cannot be satisfied. Though rework queues are considered, their stability conditions are automatically satisfied because rework arrival rate $p\bar{F}(t_i)/t_i$ is smaller than rework service rate $1/r$.

5.1 The First-Best Benchmark (Contractible Effort)

When working at capacity, the agents are continuously busy[6] and spend $t+p\bar{F}(t)r$ time units per job on average. In contrast to λ earlier, the average throughput per agent now becomes $1/(t+p\bar{F}(t)r)$. Renewal theory yields that the agents' long-run average utility rate is $b/(t+p\bar{F}(t)r)+w-a$. When effort t is contractible, the principal maximizes

$$\begin{aligned} V^{FB} &= \max_{t\geq \underline{t}, 0\leq p\leq 1, w, b} \frac{v - b - \bar{F}(t)(pc_I + (1-p)c_E) - C(p)}{t + p\bar{F}(t)r} - w \\ \text{subject to} \quad & \frac{b}{t + p\bar{F}(t)r} + w - a \geq 0 \quad \text{(IR)}. \end{aligned}$$

Since the profit rate is monotone decreasing in w and b, the IR constraint must bind and the optimization problem reduces to a two-variable problem of t and p.

[6] The rework agent in dedicated routing has idle time in steady state.

Lemma 5 *Assume an interior first-best solution $\{t^{FB}, p^{FB}\}$ exists. Then it must satisfy*

$$f(t^{FB}) = \frac{1}{p^{FB}r + \frac{1}{A(t^{FB},p^{FB})}(p^{FB}c_I + (1-p^{FB})c_E)}, \quad (4)$$

$$C'(p^{FB}) = \bar{F}(t^{FB})(c_E - c_I - A(t^{FB}, p^{FB})r), \quad (5)$$

where $A(t^{FB}, p^{FB}) = \frac{v - \bar{F}(t^{FB})(p^{FB}c_I + (1-p^{FB})c_E) - C(p^{FB})}{t^{FB} + p^{FB}\bar{F}(t^{FB})r}$.

Notice that the first-order conditions resemble Equations (1) and (2). The only difference is that the disutility of effort a is replaced with $A(t^{FB}, p^{FB})$.

5.2 Optimal Incentives for the Three Networks (Noncontractible Effort)

5.2.1 Self Routing

When effort t is not contractible, the principal maximizes

$$\begin{aligned} V^S = & \max_{0\le p\le 1, w, b} \frac{v - b - \bar{F}(t)(pc_I + (1-p)c_E)) - C(p)}{t + p\bar{F}(t)r} - w, \\ \text{subject to} \quad & \frac{b}{t + p\bar{F}(t)r} + w - a \ge 0 \quad \text{(IR)}, \\ & t \in \arg\max_{t' \ge \underline{t}} \{\frac{b}{t' + p\bar{F}(t')r} + w - a\} \quad \text{(IC)}. \end{aligned}$$

It turns out that the agents have the same optimal response as in the ample capacity case, i.e., t^S is given by equation (3). In both cases, the agents maximize their average payoff by minimizing the total expected time spent per job:

$$t^S = \arg\min_{t' \ge \underline{t}} \{t' + p\bar{F}(t')r\}$$

Doing so is optimal for the agents because the piece rate of each job is guaranteed given the opportunity of rework. Consequently, the agents' optimal effort only depends on the inspection precision p and the slope of F, thus independent of whether the agents are continuously busy or have idle time. The following lemma states the result.

Lemma 6 *With limited capacity and self routing, the agents' unique optimal effort is t^S as in Lemma 2.*

5.2.2 Dedicated Routing

Without loss of generality, we assign new jobs to agent 1 and rework to agent 2. The principal maximizes

$$\begin{aligned} V^D &= \max_{0\le p\le 1, w, b} \frac{1}{2t_1}[v - b - \bar{F}(t_1)(pc_I + (1-p)c_E)) - C(p)] - \frac{w_1 + w_2}{2}, \\ \text{subject to} \quad & \frac{(1-p\bar{F}(t_1))b}{t_1} + w_1 - a \ge 0 \quad \text{(IR1)}, \quad \frac{p\bar{F}(t_1)}{t_1}(b - ar) + w_2 \ge 0 \quad \text{(IR2)}, \\ & t_1 \in \arg\max_{t' \ge \underline{t}} \{\frac{(1-p\bar{F}(t'))b}{t'} + w_1 - a\} \quad \text{(IC1)}. \end{aligned}$$

Lemma 7 *With limited capacity and dedicated routing, the agents' unique optimal effort is*

$$t^D = \begin{cases} f^{-1}(\frac{1-p\bar{F}(t^D)}{pt^D}) & \text{if } \underline{t}f(\underline{t}) > \frac{1}{p} - 1 \\ \underline{t} & \text{if } \underline{t}f(\underline{t}) \le \frac{1}{p} - 1 \end{cases}.$$

Different from the ample capacity case, agent 1's optimal effort does not depend on b. Since agent 1 is continuously busy in the capacitated system, he does not face the trade-off between making money and having idle time. He only cares about the expected time spent per piece rate earned, and thus his successful throughput, which is given by $(1 - p\bar{F}(t))/t$.

5.2.3 Cross Routing

Unlike in the case of ample capacity, we only need to consider the queueing dynamics of the two rework queues (because the new job queues are non-empty w.p. 1 when the system is capacitated). The state of the queueing network is described by (Z_3, Z_4), where Z_{i+2} is the rework queue size for agent i. Figure 3C illustrates the state transitions of this multi-class queueing network. In steady state, $Z_3 Z_4 = 0$ because the states where both rework queues are nonempty are transient. Though we could have solved the limiting distribution for each possible state of (Z_3, Z_4) using the detailed balance equation approach, we only need the aggregate probabilities of the agents working on new jobs π_i^n and on rework π_i^r. In steady state, the queueing network must satisfy

$$\begin{aligned} \pi_i^n + \pi_i^r &= 1 \quad \text{(Law of total probability)}, \\ \lambda_{ji}\pi_j^n &= \mu^r \pi_i^r \quad \text{(Balance of agent } i's \text{ rework queue)}, \end{aligned}$$

for $i, j \in \{1, 2\}$ and $i \ne j$. Solving the equations yields

$$\pi_i^n = \frac{1 - \rho_j}{1 - \rho_i\rho_j}, \qquad \pi_i^r = \frac{\rho_j(1 - \rho_i)}{1 - \rho_i\rho_j},$$

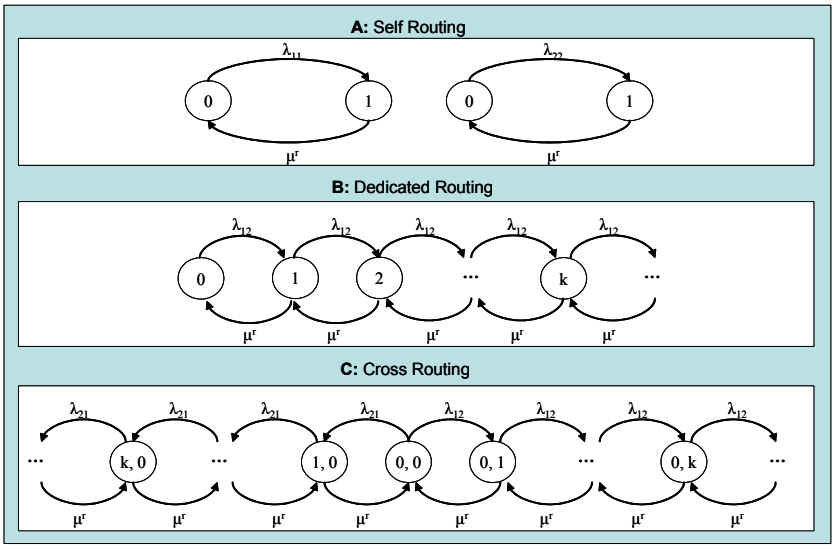

Figure 3: State Transition Diagrams of a Capacity Constrained Queueing System under the Three Rework Routing Schemes

where $\rho_i(t_i) = p\bar{F}(t_i)\frac{r}{t_i}$. Agent i's long-run average utility rate

$$U_i(t_i, t_j) = \frac{b}{1-\rho_i\rho_j}[\frac{(1-p\bar{F}(t_i))(1-\rho_j)}{t_i} + \frac{\rho_j(1-\rho_i)}{r}] + w - a.$$

Lemma 8 *With limited capacity and cross routing, if an interior symmetric Nash equilibrium* (t_i^C, t_j^C) *with* $t_i^C = t_j^C = t^C$ *exists, it must satisfy*

$$p[t^C f(t^C) + \bar{F}(t^C)][\rho(t^C)(1-\frac{r}{t^C})+1] + \rho(t^C)^2 - 1 = 0.$$

Similar to dedicated routing, the equilibrium effort only depends on p. Because we focus on symmetric Nash equilibrium, we safely suppress the subscripts distinguishing the agents. The principal's problem becomes

$$\begin{aligned} V^C &= \max_{0\le p\le 1, w, b} \frac{v - b + \bar{F}(t)(pc_I + (1-p)c_E) - C(p)}{t + p\bar{F}(t)r} - w, \\ \text{subject to} \quad & \frac{b}{t + p\bar{F}(t)r} + w - a \ge 0 \quad \text{(IR)}, \\ & t = t^C \quad \text{(IC)}. \end{aligned}$$

5.3 Comparing the Three Networks: Externality and Prisoner's Dilemma

In Section 4, we have shown that dedicated routing and cross routing impose an implicit punishment for quality failure. Here, we will highlight an additional incentive effect of cross routing: workload allocation externalities and a resulting prisoner's dilemma.

Proposition 2 *The first-best solution* $\{p^{FB}, t^{FB}\}$ *can never be achieved by the three rework routing schemes. Furthermore,* $t^S(p) < t^{FB}(p)$ *for all* $p \in (0,1)$.

The conventional self routing scheme induces lower effort than the first-best situation at any inspection precision p. As a result, self routing can never achieve first best. Dedicated routing and cross routing cannot attain first best either. Next we compare the performance of the three routing schemes in terms of effort, quality, and profit rate.

Lemma 9 *For all* $p \in (0,1)$, $t^C(p) > t^D(p) > t^S(p)$ *and therefore,* $Q^C(p) > Q^D(p) > Q^S(p)$.

Similar to the ample capacity case, self routing induces the least effort. In contrast, cross routing induces even higher effort than dedicated routing. Under cross routing, the two parallel agents impact each other in two ways: they both generate and perform rework for each other. Since rework is favorable, each agent would like the other one to send him more rework. Because rework has priority, agent i has an incentive to pass less rework to agent j so that agent j has more time to work on new jobs and pass more rework back to agent i.

Externality. The strategic interaction in the effort game results in workload allocation externalities between the agents. Whenever agent i increases effort, he not only improves his first-pass success probability, but also forces agent j to spend more time on new jobs and thus generate more rework for agent i, keeping agent $j's$ effort unchanged. Analytically,

$$\frac{\partial \pi_j^n}{\partial t_i} = -\frac{1-\rho_j}{(1-\rho_i\rho_j)^2}\frac{\partial \rho_i}{\partial t_i} > 0, \qquad \frac{\partial \pi_i^r}{\partial t_i} = -\frac{\rho_j(1-\rho_j)}{(1-\rho_i\rho_j)^2}\frac{\partial \rho_i}{\partial t_i} > 0.$$

$\partial \pi_j^n / \partial t_i > 0$ illustrates the workload externality imposed on agent j when agent i increases his first-pass effort. Since π_i^r is the fraction of time agent i spends on rework in steady state, $\partial \pi_i^r / \partial t_i > 0$ implies that agent i has more rework allocation when he increases his first-pass effort. For the same reason, agent j increases his first-pass effort to respond to agent $i's$ action. In the effort Nash equilibrium, both agents exert higher first-pass effort than under dedicated routing, resulting in better first-pass quality. Therefore, the workload allocation externalities in the effort game give cross routing superiority in inducing quality-improving effort.

Lemma 10 *For all* $p \in (0,1)$ *and at equal wage rate,*

(i) $b^C(p) > b^S(p)$;

(ii) When the wage rate is zero, $b^C(p) > b^D(p) > b^S(p)$.

Lemma 10 states that the principal pays a higher piece rate to compensate for the higher effort that agents exert under cross routing and dedicated routing. More interestingly, using this piece rate ranking, we can show that the effort equilibrium of cross routing exhibits a prisoner's dilemma.

Prisoner's Dilemma. Notice that cooperative agents would exert t^S because it minimizes the total expected time spent on each job. This cooperative outcome gives agents strictly positive utility rate because $b^C(p) > b^S(p)$, thus a better outcome for both agents than the equilibrium outcome that renders zero utility rate for both agents. Since $f(t^S) = 1/pr$,

$$\frac{\partial U_i(t_i, t^S)}{\partial t_i}|_{t_i=t^S} = \frac{b(1-\rho(t^S))}{[(1-\rho(t^S)^2)t^S]^2}[\frac{t^S}{r}(1+\rho(t^S))(\rho(t^S)(1-\frac{r}{t^S})+1)+\rho(t^S)^2-1] > 0.$$

The last inequality follows from that $\rho(t^S) < 1$ and $r \leq t^S$. Therefore, agent i has an incentive to unilaterally deviate from the cooperative outcome. (Section 6.1 elaborates on this strategic behavior and discusses incentives for collusion.) This prisoner's dilemma works in favor of the principal because it induces higher first-pass effort and thus leads to higher quality output. We now compare the principal's profit rate.

Proposition 3 $V^{FB} > \max\{V^S, V^D, V^C\}$. *The rank order of the principal's profit rate depends on the quality costs and the gross margin:*

(i) if c_I, c_E are sufficiently large or $C(\cdot)$ is sufficiently convex, $V^C > V^D > V^S$.

(ii) if v is sufficiently large, $V^S > \max\{V^D, V^C\}$.

Being capacity constrained, the principal must take into account the impact of the agents' effort on throughput. If she earns a high gross margin per job, the principal has less incentive to induce effort because higher effort than t^S leads to lower throughput and consequently lowers the revenue rate. Therefore, cross routing underperforms self routing when v is sufficiently large. However, when the costs of quality are high, it becomes critical for the principal to improve first-pass quality, making cross routing preferable to self routing.

We illustrate these effects by numerical examples. When the gross margin is high (Figure 4), there exists a threshold of c_A (c_A indicates how convex $C(\cdot)$ is), below which self routing generates the highest profit rate. In contrast, when the gross margin is low (Figure 5), cross routing always dominates the other two schemes. There exists another threshold of c_A, above which $V^C > V^D > V^S$. These results are consistent with Proposition 3.

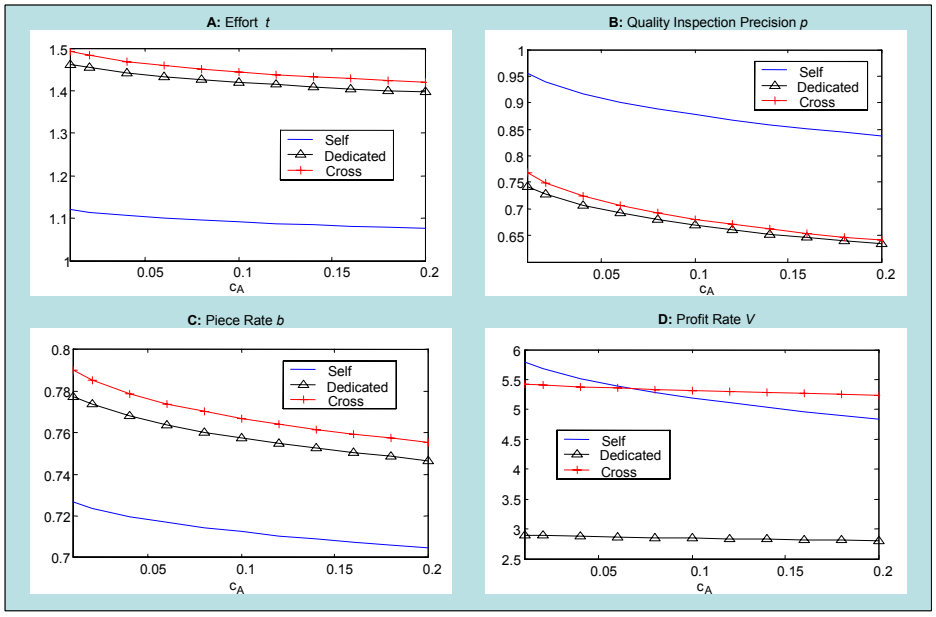

Figure 4: Financial performance depends on the appraisal cost: high gross margin ($v = 10$). $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = \frac{c_A p^2}{1-p}$, $a = 0.5$, $r = 0.5$, $w = 0$.

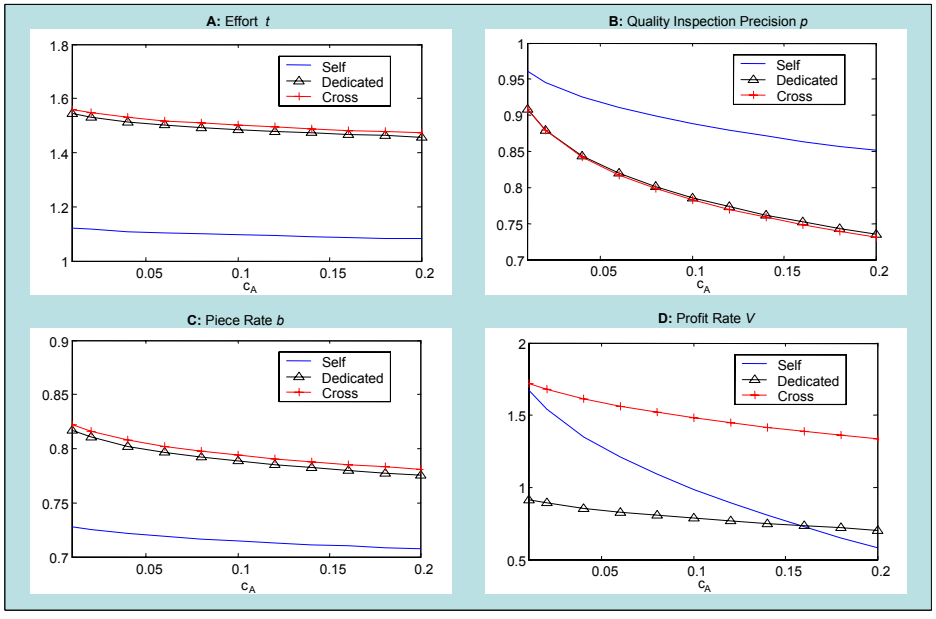

Figure 5: Financial performance depends on the appraisal cost: low gross margin ($v = 4$). $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = \frac{c_A p^2}{1-p}$, $a = 0.5$, $r = 0.5$, $w = 0$.

5.4 Equilibria with Many Agents

Similar to the case of ample capacity, the system scales up proportionally under self routing and dedicated routing and thus their incentive effects remain unchanged. In contrast, we will show that the incentive effects of cross routing reduce in an interesting way. Under cross routing, agent i's long-run average utility rate becomes

$$U_i(t_i, t_{-i}) = \frac{b}{1 + (1 - \frac{\rho_i}{N-1}) \sum_{j \neq i} \frac{\rho_j}{N-1-\rho_j}} \left[\frac{1 - p\bar{F}(t_i)}{t_i} + \frac{(1 - \frac{\rho_i}{N-1}) \sum_{j \neq i} \frac{\rho_j}{N-1-\rho_j}}{r} \right] + w - a,$$

where t_{-i} denote the vector $(t_1, ..., t_{i-1}, t_{i+1}, ..., t_N)$.

Lemma 11 *With limited capacity and cross routing, if an interior symmetric Nash equilibrium exists, the equilibrium effort t_N^C must satisfy*

$$p[t_N^C f(t_N^C) + \bar{F}(t_N^C)][\rho(t_N^C)(1 - \frac{1}{N-1}\frac{r}{t_N^C}) + 1] + \frac{1}{N-1}\rho(t_N^C)^2 - \frac{N-2}{N-1}\rho(t_N^C) - 1 = 0$$

This result generalizes the equilibrium condition of Lemma 8 to the case of N-agent.

Lemma 12 *For any $N \geq 2$ and $p \in (0, 1)$, $t_N^C(p) > t^D(p)$, where $t^D(p)$ is the optimal effort under dedicated routing as in Lemma 7.*

This lemma shows that the incentive power of cross routing is still higher than that of dedicated routing in a system with many agents. However, as the system grows larger, each agent's impact on any other specific agent diminishes. In fact, under a mild condition on the probability function $F(\cdot)$, we further establish a monotonicity property that the incentive power of cross routing decreases in the size of the system. The condition is often used in the supply chain contracting literature: *increasing generalized failure rate* (IGFR). Lariviere & Porteus (2001) define the generalized failure rate of a distribution function Φ as $\xi\phi(\xi)/\bar{\Phi}(\xi)$, where ϕ is the density of Φ and $\bar{\Phi} = 1 - \Phi$. IGFR means that $\xi\phi(\xi)/\bar{\Phi}(\xi)$ is weakly increasing for all ξ such that $\Phi(\xi) < 1$. We also need the *decreasing failure rate* (DFR) property, i.e., $\phi(\xi)/\bar{\Phi}(\xi)$ is weakly decreasing for all ξ such that $\Phi(\xi) < 1$.

Lemma 13 ***(Monotonicity).*** *Assume that $F(\cdot)$ satisfies IGFR and DFR, and that $\underline{t} f(\underline{t}) \geq 1$. For any $N \geq 2$, t_N^C is monotone decreasing in N.*

The conditions are sufficient but not necessary for the monotonicity property to hold. A canonical example is of the exponential form: $F(t) = 1 - \beta e^{-(t-\underline{t})}$. Because it has constant failure rate,

it satisfies both IGFR and DFR. Moreover, as long as $\underline{t} \geq 1/\beta$, $\underline{t}f(\underline{t}) \geq 1$ is satisfied. Though *increasing failure rate* (IFR) implies IGFR, the reverse is not true, thus making it possible for a distribution function to be both IGFR and DFR. In fact many DFR distribution functions are IGFR (Lariviere (2006)).

Notice that if we take N to the limit, the equilibrium condition reduces to $f(t) = \frac{1-p\bar{F}(t)}{pt}$, the solution of which is exactly t^D. This suggests that in the limit game, there may exist a symmetric equilibrium in which all agents play t^D. In the literature of large games, the limit of a sequence of finite games where the number of players increases to infinity can be studied as a game with continuous players in a rigorous mathematical sense (Green (1986)) or characterized using the concept of ε-equilibrium, which we adopt here.

Definition 1 ***(ε-equilibrium).*** *Let $\varepsilon \geq 0$. If $\hat{t} \in \mathbb{R}_+^N$ is an ε-equilibrium, there exists no agent i and no $t_i \in [\underline{t}, \infty)$ such that $U_i(t_i, \hat{t}_{-i}) - U_i(\hat{t}_i, \hat{t}_{-i}) > \varepsilon$.*

Notice that ε-equilibrium is a weakened notion of Nash equilibrium. If $\varepsilon = 0$, the definition reduces to that of Nash equilibrium. In words, ε-equilibrium describes the strategy profile that is within ε of the best payoff of each agent. An immediate question is why the agents would contend with something less than optimal? One interpretation of ε is that it represents the adjustment cost of discovering and using the optimal strategy (Radner (1979)). Another interpretation is bounded rationality of individual agents.

Proposition 4 *Let $\varepsilon \geq 0$. There is an N_ε such that, for all $N \geq N_\varepsilon$, there exists an ε-equilibrium in which each agent chooses t^D as in Lemma 7.*

This proposition states that in a cross routing system with many agents, it is an approximate equilibrium for all agents to behave as if rework were routed to a dedicated agent. This result is intuitive and consistent with the findings with two agents. Earlier we have shown that cross routing has higher incentives than dedicated routing because both agents can influence each other's workload of new jobs and rework in a substantial way. As the number of agents increases, the influence of each agent vanishes, which is typical of an anonymous large game. The strategic interactions thus diminish to none and the incentive power of cross routing reduces to implicit punishment, making it close to dedicated routing. The diminishing incentive power of cross routing in large systems suggests that it may be preferred for the principal to match agents into cross-routing pairs to preserve the incentives for quality.

6 Project Management Setting: Static System

Until now, we have focused on a continuous flow system where jobs are constantly assigned to agents. To check whether the strategic effects illustrated earlier extend to a static system, we now consider a project environment where a single job (i.e., project) is assigned to an agent who needs to complete it within certain period of time. This static setting also allows for a clear demonstration of the prisoner's dilemma arising from the effort game under cross routing. It is reasonable to assume that the agents maximize their utility per job in this context. For analytical purposes, we assume that there are only two possible effort levels $\{t_H, t_L\}$. With effort levels t_H and t_L, good quality output is produced with probabilities π_H and π_L, respectively. We assume $0 < \pi_L < \pi_H < 1$. As in the main model, rework takes a constant effort r, which is contractible. Further, rework is relatively less costly, specifically, $r < t_L$. Here we can allow a more general disutility[7] of effort $g(t)$, with $g(0) = 0$, $g' > 0$ and $g'' > 0$. Further more, we assume that it is optimal for the principal to induce high effort. This is the more interesting case as quality is crucial to the principal.

6.1 Self Routing

Under self routing, agent i's utility depends on his effort:

$$U_H = b - g(t_H) - p(1-\pi_H)g(r) + w, \quad U_L = b - g(t_L) - p(1-\pi_L)g(r) + w.$$

The IC constraint for inducing high effort is $U_H \geq U_L$. Equivalently,

$$p \geq \bar{p}^S = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)g(r)}.$$

To ensure high effort is implementable using the self routing scheme, we need $\bar{p}^S < 1$ and thus assume $g(r) > \frac{g(t_H)-g(t_L)}{\pi_H - \pi_L}$. This is the more interesting case because self routing would otherwise be immediately inferior. Since the throughput is limited to one unit, the principal's problem becomes minimizing the total cost per agent:

$$\begin{aligned} C^S \quad &= \quad \min_{0\leq p\leq 1, w, b} b + (1-\pi_H)(pc_I + (1-p)c_E) + C(p) + w \\ \text{subject to} \quad & \quad b - g(t_H) - p(1-\pi_H)g(r) + w \geq 0 \quad \text{(IR)}, \quad p \geq \bar{p}^S \quad \text{(IC)}. \end{aligned}$$

[7] In our main model, a linear disutility of effort is assumed because long-run analysis of the agents' utility rate requires additivity of disutility.

6.2 Dedicated Routing

To keep dedicated routing equivalent to the other two schemes, assign two projects to agent 1. Agent 1's utility depends on his effort:

$$U_H = 2[(1-p(1-\pi_H))b - g(t_H)] + w_1, \quad U_L = 2[(1-p(1-\pi_L))b - g(t_L)] + w_2.$$

The IC constraint for inducing high effort is $U_H \geq U_L$. Equivalently,

$$p \geq \bar{p}^D = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)b}.$$

The principal's problem becomes

$$\begin{aligned} C^D \quad = \quad & \min_{0\leq p\leq 1, w_1, w_2, b} b + (1-\pi_H)(pc_I + (1-p)c_E) + C(p) + \frac{w_1+w_2}{2} \\ \text{subject to} \quad & 2[\,(1-p(1-\pi_H))b - g(t_H)] + w_1 \geq 0 \quad \text{(IR1)}, \quad p \geq \bar{p}^D \quad \text{(IC1)}, \\ & 2p(1-\pi_H)(b-g(r)) + w_2 \geq 0 \quad \text{(IR2)}. \end{aligned}$$

Since $\bar{p}^D$ depends on b, high effort is always implementable as long as the piece rate b is set high enough. Similar to the ample capacity case of the main model, dedicated routing gives the principal an additional lever, i.e., b, to induce effort.

6.3 Cross Routing

Under cross routing, agent i's utility also depends on agent j's effort:

$$\begin{aligned} &U_{HH} = b - g(t_H) - p(1-\pi_H)g(r) + w, \quad U_{LH} = (1-p(\pi_H-\pi_L))b - g(t_L) - p(1-\pi_H)g(r) + w, \\ &U_{LL} = b - g(t_L) - p(1-\pi_L)g(r) + w, \quad\quad U_{HL} = (1+p(\pi_H-\pi_L))b - g(t_H) - p(1-\pi_L)g(r) + w, \end{aligned}$$

where the first and second subscripts denote agent i's and j's effort level, respectively. The IC constraint for inducing $\{H, H\}$ equilibrium outcome is $U_{HH} \geq U_{LH}$. Equivalently,

$$p \geq \bar{p}^C = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)b}$$

Similar to dedicated routing, high effort is always implementable as long as the piece rate b is set high enough. The principal's problem becomes

$$\begin{aligned} C^C \quad = \quad & \min_{0\leq p\leq 1, w, b} b + (1-\pi_H)(pc_I + (1-p)c_E) + C(p) + w \\ \text{subject to} \quad & b - g(t_H) - p(1-\pi_H)g(r) + w \geq 0 \quad \text{(IR)}, \quad p \geq \bar{p}^C \quad \text{(IC)}. \end{aligned}$$

6.4 Comparing the Three Schemes: Prisoner's Dilemma and Collusion

We first compare the lower bound on the quality inspection precision required to achieve high effort. It is simple to show that $\bar{p}^C = \bar{p}^D$ but $\bar{p}^S$ can be higher or lower. The equality between $\bar{p}^C$ and $\bar{p}^D$ implies that cross assignment of rework and assigning rework to a dedicated agent have equivalent incentive effects. When $\bar{p}^C \geq \bar{p}^S$, self routing achieves identical performance as cross routing. By contrast, when $\bar{p}^C < \bar{p}^S$, the Nash equilibrium induced between the two agents in cross routing exhibits a *prisoner's dilemma for any* $p \in (\bar{p}^C, \bar{p}^S)$. This follows from

$$\begin{aligned} U_{HL} - U_{LL} &= p(\pi_H - \pi_L)b - g(t_H) + g(t_L) > 0 \\ U_{HH} - U_{LL} &= g(t_L) - g(t_H) + pg(r)(\pi_H - \pi_L) < 0 \end{aligned}$$

In other words, even though strategy profile (t_L, t_L) gives both agents a higher payoff than the equilibrium payoff, agents will make unilateral deviation to high effort, resulting in a prisoner's dilemma. When this occurs, cross routing leads to lower costs to the principal. The next proposition states the conditions when this would occur.

Proposition 5 *$C^S \geq C^D = C^C$. The inequality is strict when c_I is sufficiently large, when c_E is sufficiently small, or when $C(\cdot)$ is sufficiently convex.*

The above results are consistent with Propositions 1 in that self routing is weakly dominated by the other two schemes. However, it differs from Proposition 3 in that c_I and c_E play different roles in determining the rank order. These differences result from that high effort is always assumed to be desirable in the current model, while the principal is allowed to choose optimal effort earlier.

Finally, we caution that the superior performance of cross routing relies on the restriction that the agents do not have future interactions. According to the Nash Reversion Folk Theorem (see Mas-Colell, Whinston & Green (1995)), a collusive outcome (t_L, t_L) can be supported with Nash reversion strategies and a sufficiently large discount factor in a repeated game. This suggests that in practice, it may be beneficial for the principal to maintain a certain level of staff turnover to prevent collusion.

7 Conclusions and Discussions

This paper investigates how incentives and judicious rework routing can improve quality and profitability of a firm using a principal-agent model integrated into a multi-class queueing network.

We demonstrate that conventional self routing of rework is always suboptimal in terms of inducing quality-improving effort. In contrast, dedicated routing and cross routing perform better in inducing effort. However, financial performance depends not only on the first-pass effort induced, but also on capacity levels, revenues, and quality costs. The more novel cross routing scheme is applicable in both manufacturing and service operations environment. The merit of this scheme lies in that the agents influence each other's workload allocation of new jobs and rework in a way that leads to higher equilibrium first-pass effort as a result of a prisoner's dilemma. This works in favor of the principal when quality is important, i.e., when quality costs are high. When the number of agents increases, the incentive power of cross routing reduces monotonically and approaches that of dedicated routing. We have also considered the effect of non-constant rework time by assuming $r = \tau - t$, where τ is a constant. The results are consistent with our main findings (see our Online Appendix for details).

We have made two methodological contributions to the agency and operations management literature. First, we study a multi-agent principal-agent model in a multi-class queueing network with endogenous queues (recall the job arrival rate of the rework queues is endogenously determined by the agents' first-pass effort). To the best of our knowledge, this is the first attempt at modeling endogenous queueing dynamics in a principal-agent framework. Second, we embed the quantity-quality trade-off in a single dimensional decision variable, i.e., the average processing time per job. This is more appropriate than the multitask principal-agent model when quantity and quality are not separable tasks of an agent's job.

We have illustrated how rework routing affects incentives in a specific queueing network formulated here. The insights of this paper will find applications in a broader network setting where a principal uses routing as an operational instrument to create incentives complementing the effects of monetary incentives. In essence, the decentralized decision making of individual agents in a queueing network creates externalities between the agents that may work in favor of the principal.

Our model has limitations. First, due to the inherent variability in queueing networks, risk aversion cannot be easily incorporated given that we conduct long-run analysis. Second, we assume that agents commit to a single first-pass effort level even though in reality they can adjust effort from time to time and thus play a dynamic game. Third, our model does not capture customer waiting costs and inventory holding costs, though they can be incorporated. When customer waiting costs are considered, pricing of the goods or services sold by the principal will depend on the agents' effort. Customer waiting also affects the principal's decision on capacity, i.e., whether to acquire

adequate staffing to provide good service or maintain high utilization of resources to minimize cost. Inventory holding costs can be incorporated straightforwardly. We believe this will change our result in one direction: the principal will have less incentives to induce effort because higher first-pass effort leads to longer flow time, and thus higher inventory holding costs. Finally, our model does not cover a few important aspects of manufacturing and service operations. From the perspective of lean operations, self routing enables quality at the source and allows the workers to learn from their own mistakes. Since the utilization of agents is not balanced in dedicated routing, cross routing may be preferred even though the two routing schemes have the same incentive effects. However, cross routing loses the specialization benefit of dedicated routing.

Acknowledgements

We thank Kevin Wilson of Memphis Auto Auction for bringing cross routing to our attention. We also greatly benefited from suggestions by Gad Allon, Sandeep Baliga, James Dana, Harry Groenelvelt, Michael Harrison, Wallace Hopp, Seyed Iravani, Johannas Horner, Ehud Kalai, Martin Lariviere.

References

Avriel, M., Diewert, W. E., Schaible, S. & Zang, I. (1988), *Generalized Concavity*, Plenum Press, New York.

Baiman, S., Fischer, P. E. & Rajan, M. V. (2000), 'Information, contracting, and quality costs', *Management Science* **46**(6), 776–789.

Benjaafar, S., Elahi, E. & Donohue, K. L. (2006), 'Outsourcing via service competition'. Working Paper.

Cachon, G. & Harker, P. T. (2002), 'Competition and outsourcing with scale economies', *Management Science* **48**(10), 1314–1333.

Cachon, G. P. & Zhang, F. (2006), 'Obtaining fast service in a queueing system via performance-based allocation of demand', *Management Science* . Forthcoming.

Gans, N., Koole, G. & Mandelbaum, A. (2003), 'Telephone call centers: Tutorial, review, and research prospects', *Manufacturing and Service Operations Management* **5**(2), 79–141.

Gilbert, S. M. & Weng, Z. K. (1998), 'Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective', *Management Science* **44**(12), 1662–1669.

Green, E. J. (1986), 'Continuum and finite-player noncooperative models of competition', *Econometrica* **52**(4), 975–993.

Gunes, E. D. & Aksin, O. Z. (2004), 'Value creation in service delivery: Relating market segmentation, incentives, and operational performance', *Manufacturing and Service Operations Management* **6**(4), 338–357.

Ha, A. Y. (2001), 'Optimal pricing that coordinates queues with customer-chosen requirements', *Management Science* **47**(7), 915–930.

Hamilton, B. H., Nickerson, J. A. & Owan, H. (2003), 'Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation', *Journal of Political Economy* **111**(3), 465–497.

Holmstrom, B. & Milgrom, P. (1991), 'Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design', *Journal of Law, Economics, and Organization* **7**, 24–52.

Juran, J. M. & Gryna, F. M. (1993), *Quality Planning and Analysis: from product development through use*, McGraw-Hill, New York.

Lariviere, M. A. (2006), 'A note on probability distributions with increasing generalized failure rates', *Operations Research* **3**(4), 293–305.

Lariviere, M. A. & Porteus, E. L. (2001), 'Selling to the newsvendor: An analysis of price-only contracts', *Manufacturing and Service Operations Management* **3**(4), 293–305.

Lazear, E. P. (2000), 'Performance and productivity', *American Economic Review* **90**(5), 1346–1361.

Mas-Colell, A., Whinston, M. D. & Green, J. R. (1995), *Microeconomic Theory*, Oxford University Press, New York.

Mendelson, H. & Whang, S. (1990), 'Optimal incentive-compatible priority pricing for the m/m/1 queue', *Operations Research* **38**(5), 870–883.

Naor, P. (1969), 'The regulation of queue size by levying tolls', *Econometrica* **37**, 15–24.

Parlakturk, A. K. & Kumar, S. (2004), 'Self-intereted routing in queueing networks', *Management Science* **50**(7), 949–966.

Plambeck, E. L. & Zenios, S. A. (2000), 'Performance-based incentives in a dynamic principal-agent model', *Manufacturing and Service Operations Management* **2**(3), 240–263.

Radner, R. (1979), 'Collusive bahavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives', *Journal of Economic Theory* **22**, 136–154.

Ren, Z. J. & Zhou, Y.-P. (2006), 'Call center outsourcing: Coordinating staffing level and service quality', *Management Science* . Forthcoming.

Reynier, D. J. & Tapiero, C. S. (1995), 'The delivery and control of quality in supplier-producer contracts', *Management Science* **41**(10), 1581–1589.

Shumsky, R. A. & Pinker, E. J. (2003), 'Gatekeepers and referrals in services', *Management Science* **49**(7), 839–856.

Van Mieghem, J. A. (2000), 'Price and service discrimination in queuing systems: Incentive compatibility of gcu scheduling', *Management Science* **46**(9), 1249–1267.

Appendix

PROOF OF LEMMA 1. Notice that when $t \geq \frac{v}{a}$, $V < 0$. Therefore, we maximize a continuous function over a compact set: $[\underline{t}, \frac{v}{a}] \times [0,1]$, implying an optimum exists. $C'(1) = \infty$ implies $p^{FB} < 1$. Since the IR constraint must bind, substitute $w = \lambda[a(t + p\bar{F}(t)r) - b]$ into the objective function and the Kuhn-Tucker conditions are

$$\begin{aligned} \frac{\partial V(t^{FB}, p^{FB})}{\partial t} &= \lambda[-a + f(t^{FB})(p^{FB}(ar + c_I) + (1 - p^{FB})c_E)] \leq 0, \\ \frac{\partial V(t^{FB}, p^{FB})}{\partial p} &= \lambda[\bar{F}(t^{FB})(c_E - c_I - ar) - C'(p^{FB})] \leq 0. \end{aligned}$$

$C'(0) = 0$ implies $p^{FB} > 0$ because $c_E - c_I > ar$ gives$\frac{\partial V(t,0)}{\partial p} > 0$. $p^{FB} \in (0,1)$ implies $(p^{FB}(ar + c_I) + (1 - p^{FB})c_E) \in (ar + c_I,\, c_E)$. When $\frac{a}{f(\underline{t})} < ar + c_I$, $\frac{\partial V(\underline{t},p)}{\partial t} > 0$ implying $t > \underline{t}$. ■

PROOF OF LEMMA 2, 3, 4. The Kuhn-Tucker condition is $-1 + prf(t) \leq 0$ with equality at interior solutions. The boundary conditions follows from that f is strictly monotone decreasing. The uniqueness follows from the second-order condition (SOC) $\lambda aprF''(t) < 0$. Proofs of Lemma 3 and 4 are similar and thus omitted. ■

PROOF OF PROPOSITION 1. From equation 1, $f(t^S) > f(t^{FB})$, implying $t^S < t^{FB}$, therefore self routing cannot achieve first best. Since $t^D = t^C = f^{-1}(\frac{a}{bp})$, set $p = p^{FB}$ and $b^* = \frac{a}{f(t^{FB})p}$ to induce t^{FB}. The rest follows from that the IR constraints are satisfied at equality. ■

PROOF OF LEMMA 1. From equation 1, $f(t^S) > f(t^{FB})$, implying $t^S(p) < t^{FB}(p) = t^D(p) = t^C(p)$. The rest follows from that Q is strictly increasing in t at any p. ■

PROOF OF LEMMA 2. Because $t^S = \arg\min_{t' \geq \underline{t}} \{t' + p\bar{F}(t')r\}$, $w^{FB}(p) = \lambda[a(t^{FB}(p) + p\bar{F}(t^{FB}(p))r) - b^{FB}(p)] = w^S(p) = \lambda[a(t^S(p) + p\bar{F}(t^S(p))r) - b^S(p)]$ implies that $b^{FB}(p) > b^S(p)$. The rest follows from $b^{FB}(p) = b^D(p) = b^C(p)$.■

PROOF OF LEMMA 5. Notice that when $t \geq \frac{v}{a}$, $V < 0$. Therefore, we maximize a continuous function over a compact set: $[\underline{t}, \frac{v}{a}] \times [0,1]$, implying an optimum exists. Assuming an interior optimum exists, optimum $\{t^{FB}, p^{FB}\}$ is then given by the first-order conditions (FOC). ■

PROOF OF LEMMA 6. Evaluating the second derivative of $U(t)$ at t^S using $f(t^S) = 1/pr$ yields that

$$U''(t^S) = \frac{bprF''(t^S)}{(t^S + p\bar{F}(t^S)r)^2} + \frac{2b(1 - prf(t^S))}{(t^S + p\bar{F}(t^S)r)^3} = \frac{bprF''(t^S)}{(t^S + p\bar{F}(t^S)r)^2} < 0.$$

Because $U(t)$ is strictly concave at any interior critical point, $U(t)$ is strictly pseudoconcave (Avriel, Diewert, Schaible & Zang (1988)) and thus t^S is a unique global maximum. The boundary conditions follows from that f is strictly monotone decreasing. ■

PROOF OF LEMMA 7. Evaluating the second derivative of $U_1(t)$ at t^D using $f\left(t^D\right) = \frac{1-p\bar{F}(t^D)}{pt^D}$ yields that

$$U_1''(t^D) = b[\frac{pF''(t^D)}{t^D} - \frac{2}{(t^D)^3}(pt^D f(t^D) + p\bar{F}(t^D) - 1)] = \frac{bpF''(t^D)}{t^D} < 0.$$

Because $U_1(t)$ is strictly concave at any interior critical point, $U_1(t)$ is strictly pseudoconcave (Avriel et al. (1988)) and thus t^D is a unique global maximum. The boundary condition follows from that $tf(t) + \bar{F}(t)$ is strictly monotone decreasing and $\bar{F}(\underline{t}) = 1$. ■

PROOF OF LEMMA 8. Assuming an interior solution, the FOC is

$$\frac{\partial U_i(t_i, t_j)}{\partial t_i} = \frac{b(1-\rho_j)}{(1-\rho_i\rho_j)^2}\frac{1}{t_i^2}[p(t_i f(t_i) + \bar{F}(t_i))(1 + \rho_j(1 - \frac{r}{t_i})) + \rho_i\rho_j - 1] = 0.$$

Let $\hat{t}_i$ be the critical point satisfying the FOC. The SOC evaluated at $\hat{t}_i$ is

$$\begin{aligned}\frac{\partial^2 U_i(\hat{t}_i, t_j)}{\partial t_i^2} &= \frac{b(1-\rho_j)}{(1-\rho_i\rho_j(\hat{t}_i))^2}\frac{1}{\hat{t}_i^2}[p\hat{t}_i F''(\hat{t}_i)(1 + \rho_j(1 - \frac{r}{\hat{t}_i})) + p(\hat{t}_i f(\hat{t}_i) + \bar{F}(\hat{t}_i))\frac{r}{\hat{t}_i^2}\rho_j + \rho_j\frac{\partial\rho_i(\hat{t}_i)}{\partial t_i}] \\ &\quad + b(1-\rho_j)\frac{\partial[\frac{1}{(1-\rho_i\rho_j)^2}\frac{1}{t_i^2}]}{\partial t_i}[p(\hat{t}_i f(\hat{t}_i) + \bar{F}(\hat{t}_i))(1 + \rho_j(1 - \frac{r}{\hat{t}_i})) + \rho_i(\hat{t}_i)\rho_j - 1].\end{aligned}$$

Substituting $\frac{\partial\rho_i(\hat{t}_i)}{\partial t_i} = -\frac{pr}{\hat{t}_i^2}[\hat{t}_i f(\hat{t}_i) + \bar{F}(\hat{t}_i)]$ and the FOC into the SOC gives

$$\frac{\partial^2 U_i(\hat{t}_i, t_j)}{\partial t_i^2} = \frac{b(1-\rho_j)}{(1-\rho_i\rho_j(\hat{t}_i))^2}\frac{1}{\hat{t}_i^2}[p\hat{t}_i F''(\hat{t}_i)(1 + \rho_j(1 - \frac{r}{\hat{t}_i}))] < 0.$$

The inequality follows from the fact that $F''(\cdot) < 0$ and that $(1 + \rho_j(1 - \frac{r}{\hat{t}_i})) > 0$. Because it is strictly concave at any interior critical point, $U_i(t_i, t_j)$ is strictly pseudoconcave in t_i (Avriel et al. (1988)), implying $\hat{t}_i$ is a unique global maximum. ■

PROOF OF PROPOSITION 2. Let $A(t,p) = \frac{v-b-\bar{F}(t)(pc_I+(1-p)c_E)-C(p)}{t+p\bar{F}(t)r}$ and $B(t,p) = t + p\bar{F}(t)r$. From equation 4, $f(t^S) > f(t^{FB})$ implying $t^S(p) < t^{FB}(p)$, therefore self routing cannot achieve first best. Claim: At any fixed p, V^{FB} is uniquely maximized at t^{FB}.

$$\frac{\partial V^2(t^{FB})}{\partial t^2} = \frac{F''(t^{FB})(pc_I + (1-p)c_E)}{B(t^{FB}, p)} + prF''(t^{FB})\frac{A(t^{FB}, p)}{B(t^{FB}, p)} < 0$$

Since $V(t)$ is strictly concave at any interior critical point, $V(t)$ is strictly pseudoconcave (Avriel et al. (1988)), and $t^{FB}(p)$ is a unique global maximum. Since the agents' optimal effort in dedicated and cross routing only depend on p and are different from the FOC of t^{FB}, first best cannot be implemented by these two schemes. ■

PROOF OF LEMMA 9. To show $t^S(p) < t^D(p)$, substituting $t^S(p)$ into the FOC of $t^D(p)$ yields that $\frac{w}{t^S(p)}[pt^S(p)f(t^S(p)) + p\bar{F}(t^S(p)) - 1] = \frac{w}{t^S(p)}[\frac{t^S(p)}{r} + p\bar{F}(t^S(p) - 1] > 0$. We show $t^D(p) < t^C(p)$ by contradiction. Suppose $t^D(p) \geq t^C(p)$ and it follows from the FOC of $t^D(p)$ that $pt^C(p)f(t^C(p)) + p\bar{F}(t^C(p)) - 1 \geq 0$. Then, $p[t^C(p)f(t^C(p)) + \bar{F}(t^C(p))][1 + (1 - \frac{r}{t^C(p)})\rho(t^C(p))] + \rho(t^C(p))^2 - 1 \geq 1 + (1 - \frac{r}{t^C(p)})\rho(t^C(p)) + \rho(t^C(p))^2 - 1 = \frac{\rho(t^C(p))}{t^C(p)}(t^C(p) + p\bar{F}(t^C(p)r - r) > 0$, contradicting the FOC of $t^C(p)$. The rest follows from that Q is strictly increasing in t at any p.■

PROOF OF LEMMA 10. Because $t^S = \arg\min_{t' \geq \underline{t}} \{t' + p\bar{F}(t')r\}$, $w^C(p) = a - \frac{b^C(p)}{t^C(p)+p\bar{F}(t^C(p))r} = w^S(p) = a - \frac{b^S(p)}{t^S(p)+p\bar{F}(t^S(p))r}$ implies that $b^C(p) > b^S(p)$. Because $w^D(p) = \frac{a[t^D(p)+p\bar{F}(t^D(p))r]-b^D(p)}{2t^D(p)}$, then zero wage rate implies $b^D(p) = a[t^D(p) + p\bar{F}(t^D(p))r]$. The inequality follows from Lemma 9. ■

PROOF OF PROPOSITION 3. First we compare the principal's profit rate at any p,

$$\begin{aligned} V^C(p) - V^S(p) &= (t^C(p) - t^S(p))\frac{(v - C(p))(pr\frac{F(t^C(p))-F(t^S(p))}{t^C(p)-t^S(p)} - 1)}{(t^C(p) + p\bar{F}(t^C(p))r)(t^S(p) + p\bar{F}(t^S(p))r)} \\ &\quad + \frac{(pc_I + (1-p)c_E)(t^C\bar{F}(t^S(p)) - t^S\bar{F}(t^C(p)))}{(t^C(p) + p\bar{F}(t^C(p))r)(t^S(p) + p\bar{F}(t^S(p))r)}. \end{aligned}$$

(i) Since $t^C(p) > t^S(p)$, it follows that $\frac{F(t^C(p))-F(t^S(p))}{t^C(p)-t^S(p)} < f(t^S(p)) = \frac{1}{pr}$ and $t^C\bar{F}(t^S(p) - t^S\bar{F}(t^C(p) > 0$. Therefore, $V^C(p) - V^S(p) > 0$ when $pc_I + (1-p)c_E$ is large enough or when $v - C(p)$ is small enough, i.e., if c_I or c_E are sufficiently large or if $C(\cdot)$ is sufficiently convex. Hence, $V^C = V^C(p^C) > V^C(p^S) > V^S(p^S) = V^S$. The first inequality follows from the optimality of V^C. (ii) $V^C(p) - V^S(p) < 0$ if v is large enough. Hence, $V^C = V^C(p^C) < V^S(p^C) < V^S(p^S) = V^S$. The last inequality follows from the optimality of V^S. Comparing V^D with V^C and V^S is similar and thus omitted. ■

PROOF OF LEMMA 11.

$$\begin{aligned} \frac{\partial U_i(t_i, t_{-i})}{\partial t_i} &= \frac{b(1-\rho_i)}{(1 + (1 - \frac{\rho_i}{N-1})\sum_{j\neq i}\frac{\rho_j}{1-\rho_j})^2}\frac{1}{t_i^2}\left\{p[t_i f(t_i) + \bar{F}(t_i)][1 + \frac{Nt_i - r}{(N-1)t_i}\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}]\right. \\ &\quad \left. -1 - (1 - \frac{\rho_i}{N-1})\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}\right\} \end{aligned}$$

Let $\hat{t}_i$ be the critical point satisfying the FOC. The SOC evaluated at $\hat{t}_i$ is

$$\begin{aligned}
\frac{\partial^2 U_i(\hat{t}_i, t_{-i})}{\partial t_i^2} &= \frac{b}{(1+(1-\frac{\rho_i}{N-1})\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j})^2}\frac{1}{\hat{t}_i^2}\left\{p\hat{t}_i F''(\hat{t}_i)[1+\frac{Nt_i-r}{(N-1)t_i}\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}]\right. \\
&\quad \left. +p[\hat{t}_i f(\hat{t}_i)+\bar{F}(\hat{t}_i)]\frac{1}{N-1}\frac{r}{\hat{t}_i^2}\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}+\frac{1}{N-1}\frac{\partial \rho_i}{\partial t_i}\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}\right\} \\
&\quad +b\frac{\partial(\frac{1}{(1+(1-\frac{\rho_i}{N-1})\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j})^2}\frac{1}{\hat{t}_i^2})}{\partial t_i}\left\{[p(\hat{t}_i f(\hat{t}_i)+\bar{F}(\hat{t}_i)][1+\frac{Nt_i-r}{(N-1)t_i}\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}]\right. \\
&\quad \left. -1-(1-\frac{\rho_i}{N-1})\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}\right\} \\
&= \frac{b}{(1+(1-\frac{\rho_i}{N-1})\sum_{j\neq i}\frac{\rho_j}{1-\rho_j})^2}\frac{1}{\hat{t}_i^2}p\hat{t}_i F''(\hat{t}_i)[1+\frac{Nt_i-r}{(N-1)t_i}\sum_{j\neq i}\frac{\rho_j}{N-1-\rho_j}]<0.
\end{aligned}$$

The last equality follows from the FOC above and that $\frac{\partial \rho_i}{\partial t_i}=-\frac{r}{t_i^2}p[t_i f(t_i)+\bar{F}(t_i)]$. The inequality follows from the fact that $F''(\cdot)<0$ and that $\rho_j<1$. Because it is strictly concave at any interior critical point, $U_i(t_i,t_j)$ is strictly pseudoconcave in t_i (Avriel et al. (1988)), implying $\hat{t}_i$ is a unique global maximum. Assuming a symmetric equilibrium gives the equation. ■

PROOF OF LEMMA 12. If $t^D(p)=\underline{t}$, we are done because $t_N^C(p)$ is interior. Now consider interior $t^D(p)$. Suppose to the contrary $t^D(p)\geq t_N^C(p)$ and it follows from the FOC of $t^D(p)$ that $p[t_N^C f(t_N^C)+\bar{F}(t_N^C)]\geq 1$ (for simplicity, we use t_N^c to denote $t_N^C(p)$.) Let $g_N(t)$ denote the FOC of the symmetric equilibrium for the N-agent system.

$$g_N(t_N^C)\geq \frac{1}{N-1}\frac{pr\bar{F}(t_N^C)}{t_N^C}[1+\frac{pr\bar{F}(t_N^C)}{t_N^C}-\frac{r}{t_N^C}]>0,$$

contradicting the optimality condition of t_N^C. ■

PROOF OF LEMMA 13. To show $t_N^C\geq t_{N+1}^C$, it suffices to show g_N decreases in both N and t. For convenience of notation, we treat N as a real number.

$$\frac{\partial g_N(t)}{\partial N}=\frac{1}{(N-1)^2}\frac{pr\bar{F}(t)(prf(t)-1)}{t}\leq 0$$

for any $t\geq t^D$ because$f(t)\leq f(t^D)\leq f(t^S)=1/pr$. Lemma 12 shows that $t^D<t_N^C$ for all $N\geq 2$. Therefore, for the range of the equilibrium solutions, g_N decreases in N. Hence $g_N(t_{N+1}^C)\geq g_{N+1}(t_{N+1}^C)=g_N(t_N^C)=0$. What is left to show is that g_N decreases in t.

$$\begin{aligned}
\frac{\partial g_N(t)}{\partial t} &= p^2 rF''(t)\bar{F}(t)(1-\frac{r}{(N-1)t})-\frac{p^2[tf(t)+\bar{F}(t)]^2 r}{t^2} \\
&\quad +ptF''(t)+\frac{pr[tf(t)+\bar{F}(t)]}{t^2}\frac{prf(t)+N-2}{N-1} \\
&\leq p^2 rF''(t)\bar{F}(t)(1-\frac{r}{(N-1)t})-\frac{p^2[tf(t)+\bar{F}(t)]^2 r}{t^2}+p[tF''(t)+\frac{tf(t)+\bar{F}(t)}{t}]
\end{aligned}$$

(b/c $f(t) \leq 1/pr$ and $r \leq t$). Since the first two terms are negative, the third term being negative is a sufficient condition for $\frac{\partial g_N(t)}{\partial t} \leq 0$, which is equivalent to $-\frac{F''(t)}{f(t)} \geq \frac{1}{t} + \frac{\bar{F}(t)}{t^2 f(t)}$. Now let us invoke the DFR assumption and from the definition of DFR we have $-\frac{F''(t)}{f(t)} \geq \frac{f(t)}{\bar{F}(t)}$. If $\frac{f(t)}{\bar{F}(t)} \geq \frac{1}{t} + \frac{\bar{F}(t)}{t^2 f(t)}$, equivalently, if $\frac{tf(t)}{\bar{F}(t)}(\frac{tf(t)}{\bar{F}(t)} - 1) \geq 0$, then the third term of $\frac{\partial g_N(t)}{\partial t}$ is negative. Satisfying the condition calls for the IGFR property and $\underline{t} f(\underline{t}) \geq 1$ so that $\frac{tf(t)}{\bar{F}(t)} \geq \frac{\underline{t} f(\underline{t})}{\bar{F}(\underline{t})} = 1$.■

PROOF OF PROPOSITION 4. Let t_i^* denote the optimal effort when all other agents choose t^D. It suffices to show $U_i(t_i^*, t_{-i}^D) - U_i(t_i^D, t_{-i}^D) \leq \varepsilon$. Claim. $t_i^* \geq t^D$. To show this, substitute t^D into the equilibrium condition given all other agents choose t^D. Because $p[t^D f(t^D) + \bar{F}(t^D)] = 1$ (If $t^D = \underline{t}$, we are done.) the condition becomes $g_N(t^D) = \frac{1}{N-1}\frac{pr\bar{F}(t^D)}{t^D}[1 + \frac{pr\bar{F}(t^D)}{t^D} - \frac{r}{t^D}] > 0$. Because the agent's problem is strictly pseudoconcave as shown in Lemma 11, $t_i^* \geq t^D$. Now

$$
\begin{aligned}
& U_i(t_i^*, t_{-i}^D) - U_i(t_i^D, t_{-i}^D) \\
= \; & \frac{b}{1 + \frac{1 - \frac{pr}{N-1}\frac{\bar{F}(t_i^*)}{t_i^*}}{1 - \frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}} \frac{pr\bar{F}(t^D)}{t^D}} [\frac{1 - p\bar{F}(t_i^*)}{t_i^*} + \frac{1 - \frac{pr}{N-1}\frac{\bar{F}(t_i^*)}{t_i^*}}{1 - \frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}} \frac{p\bar{F}(t^D)}{t^D}] - \frac{b}{t^D + pr\bar{F}(t^D)} \\
\leq \; & \frac{b}{1 + \frac{p\bar{F}(t^D)}{t^D}} [\frac{1 - p\bar{F}(t^D)}{t^D} + \frac{1}{1 - \frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}} \frac{p\bar{F}(t^D)}{t^D}] - \frac{b}{t^D + pr\bar{F}(t^D)}
\end{aligned}
$$

because $1 - \frac{pr}{N-1}\frac{\bar{F}(t_i^*)}{t_i^*} \geq 1 - \frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}$ and $\frac{1-p\bar{F}(t_i^*)}{t_i^*} \leq \frac{1-p\bar{F}(t^D)}{t^D}$. Choose N_1 large enough s.t. $\frac{pr}{N_1 - 1}\frac{\bar{F}(t^D)}{t^D} \leq \frac{1}{2}$. Then $\frac{1}{1 - \frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}} \leq 1 + 2\frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}$. It follows that

$$
\begin{aligned}
LHS \; & \leq \; \frac{b}{1 + \frac{p\bar{F}(t^D)}{t^D}} [\frac{1 - p\bar{F}(t^D)}{t^D} + \frac{p\bar{F}(t^D)}{t^D} + 2\frac{pr}{N-1}\frac{\bar{F}(t^D)}{t^D}\frac{p\bar{F}(t^D)}{t^D}] - \frac{b}{t^D + pr\bar{F}(t^D)} \\
& = \; \frac{1}{N-1}\frac{2rb(\frac{p\bar{F}(t^D)}{t^D})^2}{1 + \frac{p\bar{F}(t^D)}{t^D}}.
\end{aligned}
$$

Now choose N_2 large enough s.t. LHS $\leq \varepsilon$. Let $N_\varepsilon = \max(N_1, N_2)$. ■

PROOF OF PROPOSITION 5. Since all the IR constraints bind, the principal's problems under dedicated and cross routing are identical, implying $C^D = C^C$. The principal's problem under cross routing is equivalent to

$$\min_{0 \leq p \leq 1} g(t_H) + p(1 - \pi_H)g(r) + (1 - \pi_H)(pc_I + (1-p)c_E) + C(p)$$

The FOC is $C'(\hat{p}) = (1 - \pi_H)(c_E - c_I - g(r))$. Then, $b = \frac{g(t_H) - g(t_L)}{(\pi_H - \pi_L)p}$ and $w = g(t_H) + p(1 - \pi_H)g(r) - b$. The principal's problem under self routing is equivalent to

$$
\begin{aligned}
& \min_{0 \leq p \leq 1, w, b} g(t_H) + p(1 - \pi_H)g(r) + (1 - \pi_H)(pc_I + (1-p)c_E) + C(p) \\
& \text{s.t. } p \geq \bar{p}^S.
\end{aligned}
$$

Then w and b are determined by $w = g(t_H) + p(1-\pi_H)g(r) - b$.If $\bar{p}^S \leq \hat{p}$, then $C^S = C^C$. If $\bar{p}^S > \hat{p}$, then $C^S > C^C$. This occurs when c_I is sufficiently large, when c_E is sufficiently small, or when $C(\cdot)$ is sufficiently convex, as evident from the FOC of $\hat{p}$. ■

Online Appendix

Dependent Rework Time

We now relax the assumption that the rework time has a constant mean r. We let r depend on the first-pass effort t, i.e., $r = \tau - t$, where τ is a constant. In this subsection, we will delve directly into the results without laying out the optimization problems of the principal and the agents. Note that these problems are identical to the ones presented earlier except that the rework time r is replaced with $\tau - t$ wherever appropriate. We summarize the agents' optimal effort in Table 2.

With ample capacity, we show that dedicated routing and cross routing can implement the first-best effort and quality inspection precision and enables the principal to achieve the first-best profit rate, while self routing cannot. With limited capacity, while the agents' problems remain well behaved, the comparison of the three rework routing schemes becomes analytically less tractable. To verify that our main results presented earlier still hold with dependent rework time, we conducted a numerical study. Our numerical results are consistent with the main results earlier. When the gross margin is high, Figure 6 shows that there exists a threshold of c_A below which self routing leads to highest profit rate. In contrast, when the gross margin is low, Figure 7 shows that cross routing always dominates the other two routing schemes. These numerical findings are consistent with results in Proposition 3.

	Ample Capacity	Limited Capacity
Self	$f^{-1}(\frac{1-p\bar{F}(t^S)}{p(\tau-t^S)})$	$f^{-1}(\frac{1-p\bar{F}(t^S)}{p(\tau-t^S)})$
Dedicated	$f^{-1}(\frac{a}{pb})$	$f^{-1}(\frac{1-p\bar{F}(t^D)}{pt^D})$
Cross	$f^{-1}(\frac{a}{pb})$	$\tau p\rho(t^C)[t^C f(t^C) + \bar{F}(t^C) - \frac{(t^C)^2}{\tau}][\frac{1}{\tau-t^C} - \frac{1-p\bar{F}(t^C)}{t^C}]$ $-(1-\rho(t^C)^2)[1 - pt^C f(t^C) - p\bar{F}(t^C)] = 0$

Table 2: The agents' optimal effort under dependent rework time (assuming interior solutions)

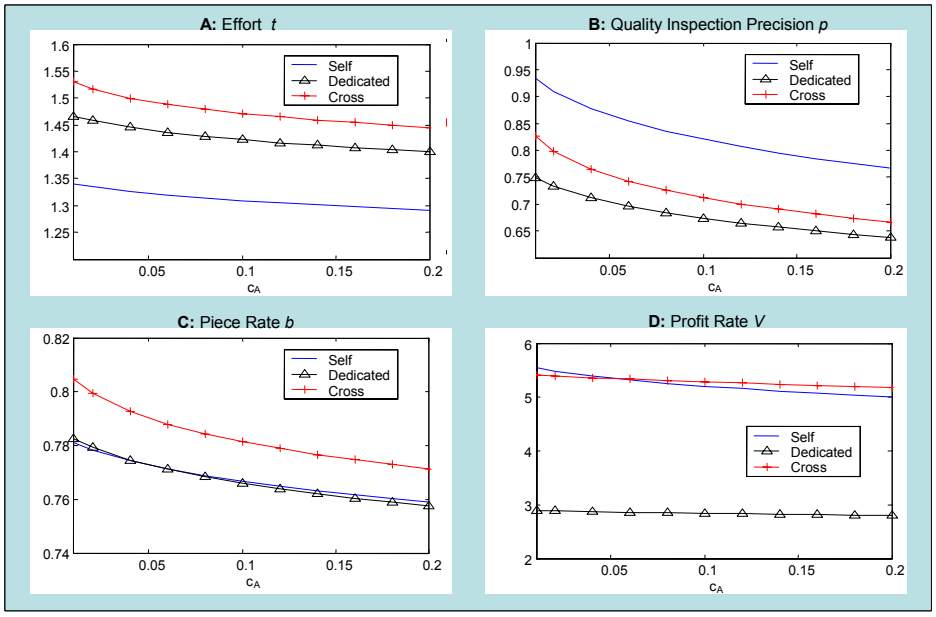

Figure 6: Equilibrium performance depends on the appraisal cost: high gross margin ($v = 10$). $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = \frac{c_A p^2}{1-p}$, $\tau = 2$, $a = 0.5$, $w = 0$.

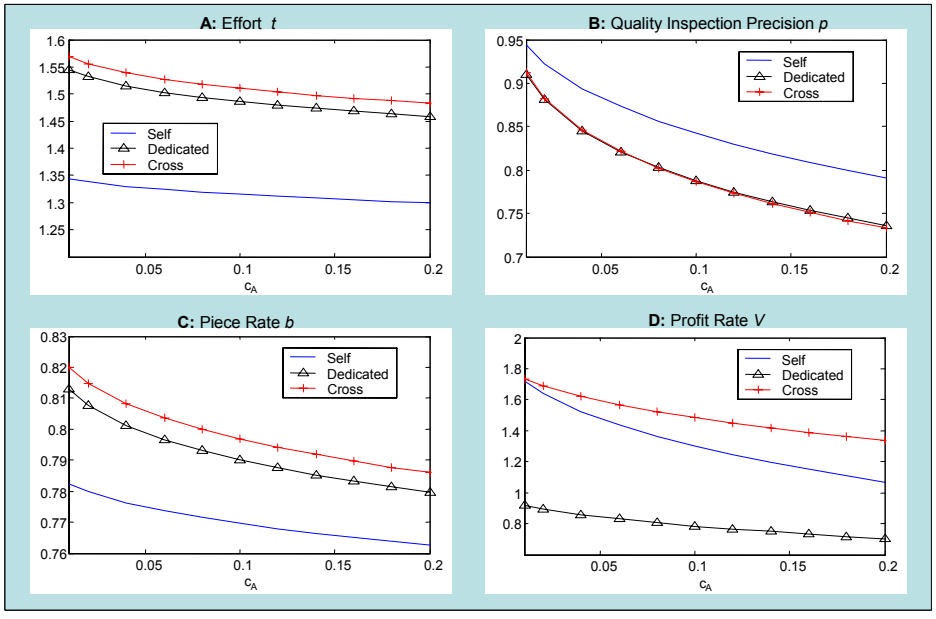

Figure 7: Equilibrium performance depends on the appraisal cost: low gross margin ($v = 4$). $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = \frac{c_A p^2}{1-p}$, $\tau = 2$, $a = 0.5$, $w = 0$.