M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 2, Spring 2009, pp. 254–273
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1102 | 0254

informs®

DOI 10.1287/msom.1080.0212
© 2009 INFORMS

Incentives for Quality Through Endogenous Routing

Lauren Xiaoyuan Lu
Kenan-Flagler Business School, University of North Carolina, Chapel Hill, North Carolina 27599,
lauren_lu@unc.edu

Jan A. Van Mieghem, R. Canan Savaskan
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
{vanmieghem@northwestern.edu, r-savaskan@kellogg.northwestern.edu}

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 multiclass queueing network. We show that conventional self-routing of rework cannot 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 the greatest incentives for quality. Firm profitability depends on demand levels, revenues, and quality costs. When the number of agents increases, the incentive effect of cross-routing reduces monotonically and approaches that of dedicated routing.

Key words: queueing networks; routing; Nash equilibrium; quality control; piece rate; epsilon equilibrium
History: Received: July 14, 2006; accepted: December 16, 2007. Published online in *Articles in Advance* June 4, 2008.

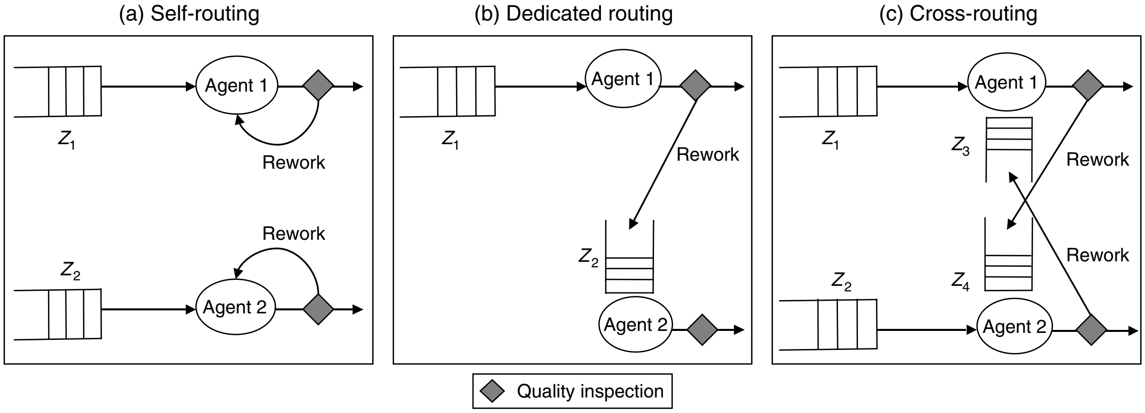
1. Introduction

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 Memphis Auto Auction, a wholesale automotive liquidator of used vehicles that employs two teams of employees to clean and detail vehicles in parallel. The employees are paid piece rates only if their jobs pass quality inspection; the quality control leader is paid a salary plus a bonus based on overall work quality. The firm ties compensation to quality through an unconventional rework-routing scheme illustrated in Figure 1(c) that we call *cross-routing*. This cross-routing of rework contrasts with the two traditional practices that assign rework back to the team that generates the defect (Figure 1(a)) or to a dedicated rework team (Figure 1(b)) and that 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 understand how and why these three routing schemes may differ in quality incentives and firm profits. Our analysis uses a principal-agent model with endogenous piece rate, quality control, and routing in a multiclass queueing network. Rework routing impacts agent incentives to exert quality-improving effort in two 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

Figure 1 **Three Rework-Routing Schemes**



of a noncooperative effort game played between 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. When the demand for finished goods or services is unlimited, 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. This benefits Agent 1 but reduces Agent 2's pay. To counteract this negative externality from Agent 1's action, Agent 2 increases first-pass effort. Consequently, both agents exert high effort and receive low rework allocation in equilibrium. This equilibrium is 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 the cooperative outcome if both exerted low effort. 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."

A higher first-pass effort produces fewer internal and external defects, but it does not always lead to higher profits for the principal. On the 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 and Gryna 1993): internal failure costs, external failure costs, and appraisal costs. On the other hand, excessively high first-pass effort lowers effective throughput, as the agents spend more effort (processing time) per job.[1] Because quality incentives can be deemed as a form of prevention costs, our model covers all 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 nonconformance with appraisal and prevention costs. Our model adds an additional dimension to this cost minimization view of quality management—throughput and thus revenues also impact a firm's quality control policies.

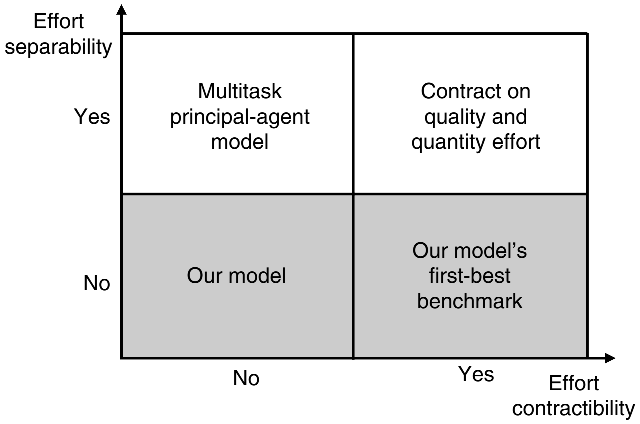
From an economic perspective, it is instructive to highlight and discuss two important assumptions in our model: contracting variables and contracting instruments.

1.1. Contracting Variables

In our model, effort is the average first-pass service time chosen by the agents. It determines both quality and quantity (throughput) but cannot be directly contracted on. To put these restrictions in perspective, consider the effort contractibility and separability matrix of Figure 2. According to Holmstrom and

[1] In our model, high first-pass effort lowers gross throughput. However, effective throughput (i.e., throughput passing quality inspection) is concave in first-pass effort.

Figure 2 **Effort Contractibility and Separability Matrix**



	Effort contractibility	
Effort separability	No	Yes
Yes	Multitask principal-agent model	Contract on quality and quantity effort
No	Our model	Our model's first-best benchmark

Milgrom (1991), effort separability means that a job involves multiple tasks that can be carried out by different agents at different times. When quality and quantity efforts can be separated and contracted on, incentive problems disappear. When they are separable but not contractible, the celebrated multitask principal-agent model of Holmstrom and Milgrom (1991) provides an optimal mechanism. In our model, quality and quantity efforts are assumed inseparable so that the intrinsic trade-off between them is modeled with a single-dimensional variable. This inseparability condition has the benefit of highlighting the essential quality-quantity trade-off in the simplest manner: we assume that spending more time per job increases quality (and obviously reduces gross throughput). Admittedly, this modeling simplification comes at a cost: working slowly may also reflect shirking that may not be observable to the firm. A more realistic approach would combine our model with a multitask model by separating true quality-enhancing time from shirking time. However, the effort dynamics that we want to study (namely, agents substitute quantity for quality by working fast) can be captured without such an extension.

In a long-run repeated setting, the average service time may become observable to the principal. Nevertheless, we assume it cannot be directly contracted on. If it could, first-best outcomes could be achieved and would give the theoretical benchmark for our model. However, in real life, relying solely on an average service-time contract may not ensure good quality. According to Gans et al. (2003), an incentive scheme that rewards agents in a call center for maintaining low average handle time can lead them to hang up on customers. Often the "ideal" service time needed to complete each job is highly variable—certain jobs take longer than others. In a service setting, employees usually have closer contacts with customers and thus can be in a better position than their managers to make the right service decisions. Contracting on service time directly does not bring such an advantage. Moreover, as mentioned earlier, real-life effort includes both the time taken and whether that time is spent on quality improvement or on shirking, which may be unobservable. Contacting on service time would require management to closely monitor employees, which could be costly to the firm.

1.2. Contracting Instruments

In our model, we restrict the principal's contracting instruments to wage, piece rate, and routing control. There are multiple other methods for achieving or approaching the first-best solution. For example, the implicit punishment imposed by assigning rework to another agent can be made explicit through a monetary punishment whenever a defect is identified. Similarly, a bonus can be paid whenever a new job passes quality inspection. These contracts are frequently used in practice. The online appendix shows that they both achieve first-best outcomes. Rather than including penalties or bonuses, which have already been extensively studied in the literature, we study contracts that only contain wage and piece-rate components and add rework routing as an operational instrument. Thus, our approach marries economics with operations strategy and processing network design.

Cross-routing of rework not only offers an alternative to penalty or bonus contracts, but it also bears some potential benefits. In contrast to a penalty contract, with cross-routing, the principal always pays the full piece rate per job. This prevents the principal's moral hazard of raising the quality standard or cheating on quality inspection to overly penalize the agents and may thus be perceived more positively by the agents. In a call center with many agents, cross-routing has the additional benefit of efficiency, because unresolved issues can be routed to any idle agent rather than to the first-contact agent, who might be busy at the time of customer call-back.[2]

Using an analytical model, we establish the following results for the three schemes of Figure 1. Traditional self-routing of rework cannot 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 demand levels, revenues, and quality costs. Under limited demand, dedicated routing and cross-routing both achieve the first-best profit rate, but self-routing does not. Under unlimited demand, cross-routing generates the highest profit rate when appraisal, internal failure, or external failure costs are high, and 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. Using the concept of epsilon equilibrium, we show that the incentive effect of cross-routing approaches that of dedicated routing as the number of agents increases to infinity.

The remainder of the paper is organized as follows. Section 2 reviews related literature, and §3 lays out the main model. Sections 4 and 5 analyze the networks under limited demand and unlimited demand, respectively. In both of the latter two sections, we first derive the first-best benchmark, then analyze the three rework-routing schemes and finally compare their performance. Section 6 examines an extension with dependent rework time, and §7 concludes. All proofs are included in the online appendix. 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 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 and Milgrom (1991) explain the trade-offs between inducing effort on quantity versus 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 per job. Moreover, we provide theoretical support for the notion that smart routing of rework can induce quality-improving effort even when using 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 exert system suboptimal quality effort under self-routing.

Holmstrom and 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 another. 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, and rework is assumed to have no uncertainty in quality. Supporting Holmstrom and Milgrom's theory that tasks should be separated according to their measurability characteristics, we show that dedicated routing provides higher-quality incentives than self-routing.

The second relevant stream of literature is on the economics of quality control and inspection in a game-theoretic setting. Papers in this stream mostly consider quality-related contractual issues between firms and are only tangentially related to our work. For example, Reynier and Tapiero (1995) study the

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

effect of contract parameters and warranty costs on the choice of quality by a supplier and the quality control policy by a buyer. Baiman et al. (2000) focus on how the contractibility of quality-related information impacts the product quality and profits of a supplier and a buyer. Our work studies how rework routing and the costs of quality affect 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 postcontractual 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 systems with unlimited demand, we allow agents' effort level (i.e., processing time) to directly impact system throughput. 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 the social optimum and prevent performance degradation as a result of customers' self-interested behavior. Many followers (e.g., Mendelson and Whang 1990, Van Mieghem 2000, Ha 2001) also design pricing mechanisms to achieve system optimal performance, but none of these works models quality inspection and rework.

Principal-agent models in queueing systems have been explored in the operations management literature. Plambeck and Zenios (2000) study incentives in a dynamic setting, where an agent's effort is not observed by the principal but influences the transition probabilities of a system. Similarly in our model, probabilistic routing is determined by agents' effort. Hasija et al. (2008) also assume an unobservable production rate to study outsourcing of call center service when agents have different productivity types. Ren and Zhou (2008) use a multitask principal-agent model to study the coordination of capacity and service quality decisions in call center outsourcing. Gunes and 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. In contrast to these papers, which focus on single-agent problems, our model captures strategic interactions between multiple agents. Parlakturk and Kumar (2004) construct a multiagent model to study how customers' self-interested routing behavior impacts system performance. Unlike the agents in our model, the customers do not exert effort or act as servers and thus are not paid or employed by the principal.

Our paper is closely related to the paper by Shumsky and Pinker (2003) in that the principal designs incentives to induce effort in steady state, but it differs in two important ways: first, we explicitly model the queueing network dynamics and consider the case where effort impacts system throughput. Second, the principal in our model hires multiple agents whose expected utility rates are interdependent. Therefore, we need to investigate agents' strategic interactions and derive the effort Nash equilibrium. The novelty of our model also lies in that we model endogenous queues—the arrival rate of rework is endogenously determined by the agents' effort.

Sharing a common theme with Gilbert and Weng (1998), Benjaafar et al. (2007), and Cachon and Zhang (2007), our paper uses an operational instrument to create incentives in a multiagent processing system. In all these papers, service rate choice may be observable to the principal but is not directly contracted on. 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).

Multiagent games in queueing systems can also be found in Cachon and Harker (2002), whose model, however, does not involve a principal. A novel feature of our queueing game-theoretical modeling is borrowing the concept of epsilon equilibrium from the literature of large games. To the best of our knowledge, this is one of the first papers in the operations management literature to use epsilon equilibrium to study the strategic behavior of agents in a large operational system.

In our Memphis Auto Auction example, the firm employs teams to complete jobs. In this paper, we will treat teams as agents and ignore the intrateam incentive issues that may arise because of free riding and collaboration. A relevant reference for team incentives is Hamilton et al. (2003); that study empirically investigates the impact of teams on productivity. It distinguishes the individual piece rate used in flow production from the group piece rate used in modular production and finds that the group piece rate has a stronger incentive effect on productivity than the individual piece rate because of collaboration among team members.

3. Model

3.1. 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 the sampling frequency of inspection. We assume that the principal commits to p once announced.

Each agent chooses first-pass effort (average 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. This assumption is not unrealistic—a fixed wage may still elicit some effort because not all work is unpleasant to the agents, as argued by Holmstrom and Milgrom (1991). We assume that each agent's service times are independent and identically distributed with a finite mean t, which captures the randomness in service time. This strategic decision variable drives the output quality. We assume that the agents adopt open-loop strategies and thus the average 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 returns to effort, and denote $f = F'$ and $\bar{F} = 1 - F$. On identifying defects, the principal routes the rework 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 the demand environment of the system. Under limited demand, 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 the exogenous demand arrival rate. Such an operation resembles a make-to-order system operated in a "pull" mode. In contrast, when the system has unlimited demand, the agents are continuously busy and their effort levels directly impact throughput, or the true capacity of the system. Such an operation mimics a production system operated in a "push" mode with an unlimited supply of raw materials and unlimited demand for finished goods or services.[3] This type of operation is often studied in the literature of queueing production systems (see Conway et al. 1988, Spearman and Zazanis 1992, Van Oyen et al. 2001, and Kekre et al. 2003).

The rework service times are independent and identically distributed with a mean of r, where r is a constant and common knowledge. Because defects have to be corrected as instructed by the principal, we assume that rework effort is contractible; i.e., there is no moral hazard problem in rework. We argue that even if agents may exhibit opportunistic behavior in rework, the effect is limited because identified defects have to be corrected completely. Furthermore, rework has nonpreemptive priority over new jobs. This priority rule is adopted because under unlimited demand, agents can always be engaged in new jobs. Without the priority rule, defects may never be reworked. Finally, we assume that rework takes less time than the minimum first-pass effort:

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

[3] One likely environment for such a system is a firm that keeps processing capacity constantly busy to meet the baseline demand while using extra capacity to handle the variable demand.

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 §4.3. This assumption relates our model to the type of environment where rework is quick and relatively inexpensive. A likely environment is one where many possible sources of defects exist and quality inspection requires specialized techniques and diagnosis. Once the source of error is identified, remediation of the defect is simple. For example, in a tax service firm an error in tax preparation may occur if a tax code is not properly followed or not all information is entered correctly. Identifying these errors may require extra expertise and patience. But once an error is identified, fixing it is relatively straightforward.

3.2. Financial Flows and Incentives

Under self-routing and cross-routing, 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, Agent 1 earns wage rate w_1 and piece rate b when completing a new job that passes quality inspection. Agent 2 is paid by a wage rate w_2 and piece rate b when completing a rework. The average of w_1 and w_2 is the equivalent wage rate for comparison purposes.

The agents' disutility of effort per unit time is a. Without loss of generality, we normalize the agents' reservation utility to zero. The principal earns gross margin v per completed job that passes quality inspection, pays agents, and incurs three quality costs fclassified as in Juran's cost-of-quality framework as follows: (1) $C(p)$ is the appraisal cost per new job. 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) c_I is the internal failure cost per new job. (3) c_E is the external failure cost per new job. (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 Under Limited Demand

We assume that the principal maintains a sufficient staffing level to complete all jobs with an appropriate waiting 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 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. 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. The job arrival rate is λ per agent, so 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

$$V^{FB} = \max_{0\le p\le 1,\, t\ge \underline{t},\, w,\, b} \quad \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E) - C(p)] - w,$$

$$\text{subject to} \quad \lambda[b - a(t + p\bar{F}(t)r)] + w \ge 0 \qquad \text{(IR)}.$$

The individual rationality (IR) constraint specifies the agents' outside option. Note that the IR constraint is identical for both agents. Because 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 one of two variables: t and p. Let $\{t^{FB}, p^{FB}\}$ denote the first-best solution.[4] Because

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

$\rho_i = \lambda(t^{FB} + p\bar{F}(t^{FB})r)$, the stability condition becomes $\lambda < 1/(t^{FB} + p\bar{F}(t^{FB})r)$. (Only for specific instances of F can this inequality be solved explicitly in terms of model primitives.) For a stable system, Lemma 1 characterizes the first-best solution.

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

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

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

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" $F(\cdot)$, whose derivative measures the marginal return of effort. It is simple to show that $\partial^2 V/\partial t \partial p < 0$; i.e., t and p are strategic substitutes. The principal can select from infinite pairs of wage rate and piece rate (w^{FB}, b^{FB}) to satisfy the IR constraint at equality. The principal is the Stackelberg leader and the agents earn zero utility rate in equilibrium, so the principal's objective is identical to a central planner's. Therefore, the first-best solution achieves the Pareto optimum for the entire system. From now on, we will implicitly assume that the conditions in Lemma 1 are satisfied so that an interior first-best solution exists.

4.2. Optimal Incentives for the Three Networks (Noncontractible Effort)

4.2.1. Self-Routing. When effort t is not contractible and rework is routed back to the originating agent, the principal maximizes

$$V^S = \max_{0 \le p \le 1,\, w,\, b} \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E) - C(p)] - w,$$

$$\text{subject to } \lambda[b - a(t + p\bar{F}(t)r)] + w \ge 0 \quad \text{(IR)},$$

$$t \in \arg\max_{t' \ge \underline{t}} \lambda[b - a(t' + p\bar{F}(t')r)] + w \quad \text{(IC)}.$$

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

Lemma 2. *Under limited demand and self-routing, the agents' unique optimal effort is*

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

Because the agents have sufficient time to complete all jobs and always earn the piece rate for each job, their 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 $\lambda < 1/(t^S + p\bar{F}(t^S)r)$.

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 job arrival rate 2λ. The principal maximizes

$$V^D = \max_{0 \le p \le 1,\, w,\, b} \lambda[v - b - \bar{F}(t)(pc_I + (1-p)c_E) - C(p)] - \frac{w_1 + w_2}{2},$$

subject to

$$2\lambda[(1 - p\bar{F}(t))b - at] + w_1 \ge 0 \quad \text{(IR1)},$$

$$2\lambda p\bar{F}(t)(b - ar) + w_2 \ge 0 \quad \text{(IR2)},$$

$$t \in \arg\max_{t' \ge \underline{t}} 2\lambda[(1 - p\bar{F}(t'))b - at'] + w_1 \quad \text{(IC1)}.$$

Because Agent 2's rework effort is contractible, only Agent 1's IC constraint is needed.

Lemma 3. *Under limited demand and dedicated routing, Agent* 1*'s unique optimal effort is*

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

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's 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 $\lambda < \min\{1/2t^D, 1/2p\bar{F}(t^D)r\} = 1/2t^D$.

4.2.3. Cross-Routing. When rework is assigned to the parallel agent, a rework queue is generated whose 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 multiclass queueing network illustrated in Figure 1(c), we define the following rates for $i, j \in \{1, 2\}$, and $i \neq j$:

- Agent i's new job service rate: $\mu_i^n = 1/t_i$
- Agent i's defect generation rate: $\lambda_{ij} = p\bar{F}(t_i)/t_i$
- Rework service rate: $\mu^r = 1/r$ (same for both agents).

Let a four-dimensional vector (Z_1, Z_2, Z_3, Z_4) represent the state of the four queues in the system (two new job queues and two rework queues, as shown in Figure 1(c)). For general interarrival and service time distributions, queue stationary distributions cannot be solved in analytical closed form without approximation. (One could compute them for Poisson arrivals and exponential service times, which is a special case of our model.) To compute the long-run average payoffs, however, it suffices to know the stationary average fractions of time that agent i is idle, works on new jobs, and works on rework, are denoted by π_i^0, π_i^n, and π_i^r respectively. These three fractions must add up to 1 and also satisfy the law of flow conservation:

$$\pi_i^0 + \pi_i^n + \pi_i^r = 1,$$
$$\lambda = \mu_i^n \pi_i^n,$$
$$\lambda_{ji}\pi_j^n = \mu^r \pi_i^r,$$

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 and the second term is his average reward rate from completing rework generated by agent j.

LEMMA 4. *Under limited demand and cross-routing, the unique Nash equilibrium of the agents' effort game is* (t^C, t^C), *where* $t^C = t^D$, *as defined in Lemma* 3.

Surprisingly, each agent's optimal effort in equilibrium is independent of the 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 incentive effect 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. Because agent i's utilization $\rho_i = \lambda(t_i + p\bar{F}(t_j)r)$, the stability condition becomes $\lambda < 1/(t^C + p\bar{F}(t^C)r)$. The principal maximizes

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

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 pronounced. 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 a 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 because the reduction in external failure costs cannot compensate for the high costs of internal repair.

Proposition 1 (Limited Demand). *The principal cannot attain the first best using self-routing. In contrast, she can attain the first best using dedicated routing with contract* $\{p^{FB}, w_1^*, w_2^*, b^*\}$ *or cross-routing with contract* $\{p^{FB}, w^*, b^*\}$, *where* $b^* = ar + c_I + ((1-p^{FB})/p^{FB})c_E$, $w_1^* = 2\lambda[at^{FB} - (1-p^{FB}\bar{F}(t^{FB}))b^*]$, $w_2^* = 2\lambda p^{FB}\bar{F}(t^{FB})(ar-b^*)$, *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, the principal can attain the first best using either dedicated routing or cross-routing. The optimal piece rate b^* is chosen to induce the first-best effort, and the wage rates are chosen to meet the agents' reservation utility. Notice that the optimal wage rates can be negative. However, this should not be taken literally. The overall payment from the principal to the agents is always positive. Moreover, recall that the reservation utility in our model is normalized to zero for notational simplicity. If instead a positive reservation utility is used, the optimal wage rates would increase and possibly become positive.

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 under the first-best contracts, the principal achieves the highest possible profit rate under these two routing schemes. We further compare the three schemes using first-pass effort, quality output, and piece rate. Because they are impacted not only by the routing policy but also by the quality inspection precision p, we need to hold p constant when comparing the three schemes. We will use $t(p)$, $Q(p)$, and $b(p)$ to denote the optimal effort, quality output, and piece rate at any given p.[5]

Table 1 **Comparing First-Pass Efforts, Piece Rates, and Profit Rates of the Three Schemes**

	Limited demand			Unlimited demand		
Routing	Effort	Piece rate	Profit rate	Effort	Piece rate	Profit rate
Self	Low	Low	Low	Low	Low	Depends on quality costs and gross margins
Dedicated	High	High	High	Medium	Medium	
Cross	High	High	High	High	High	

Corollary 1 (Limited Demand). $t^{FB}(p) = t^D(p) = t^C(p) > t^S(p)$ *and* $Q^{FB}(p) = Q^D(p) = Q^C(p) > Q^S(p)$.

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 the fact that the agents lose the effort spent on the jobs that fail quality inspection.

Corollary 2 (Limited Demand). *At any given wage rate,* $b^{FB}(p) = b^D(p) = b^C(p) > b^S(p)$.

Because there exist infinite pairs of w and b that satisfy the IR constraint at equality for the first-best solution, we need to fix w to have a meaningful comparison of b. 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 fail inspection. A summary of the comparison is displayed in the left column of Table 1.

4.4. Equilibria with Many Agents

So far we have considered only an operation with two agents. It is interesting to see if the results hold in a large operation with many agents. Let N, a positive integer, 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 policy is. For dedicated routing, we treat each pair of agents as the basic unit, and thus 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 per agent. For cross-routing, we treat the agents as interchangeable and route rework

[5] It is meaningful to define these functions. Under self-routing, t^S only depends on p. Under dedicated routing, it is optimal to set $f(t)$ according to Equation (1) at any given p. Thus, the optimal piece rate is $ar + c_I + ((1-p)/p)c_E$, a function of p. Therefore, $t^D(p, b) = t^D(p, b(p)) = t^D(p)$. The same reasoning applies to cross-routing.

generated by agent i to all the other agents with equal probability. The long-run utility rate of agent i is

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

where t_{-i} denotes the vector $(t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_N)$. It follows immediately that $t_i^C = f^{-1}(a/pb)$, which is identical to that of the two-agent system. In summary, under limited demand, the system can be scaled up proportionally, and the incentive effects of the three routing schemes remain unchanged.

5. Incentives and Routing Under Unlimited Demand

In contrast to the limited demand case, the throughput of the system under unlimited demand is endogenous and depends on the agents' efforts. 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 against 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 is sufficient such that the stability conditions of new job queues (described in the previous section) cannot be satisfied. However, the stability conditions of rework queues are automatically satisfied because rework arrival rate $\pi_i^n p\bar{F}(t_i)/t_i$ is smaller than rework service rate $1/r$.

5.1. First-Best Benchmark (Contractible Effort)

When demand is unlimited, the agents are continuously busy[6] and spend $t + p\bar{F}(t)r$ time units per job on average. In contrast to the limited demand case, the effective 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

$$V^{FB} = \max_{t\geq \underline{t},\, 0\leq p\leq 1,\, w,\, b} \frac{v-b-\bar{F}(t)(pc_I+(1-p)c_E)-C(p)}{t+p\bar{F}(t)r} - w$$

$$\text{subject to } \frac{b}{t+p\bar{F}(t)r}+w-a\geq 0 \quad \text{(IR)}.$$

Because 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.

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

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

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

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 $A(t^{FB}, p^{FB})$. For the rest of §5, we assume that an interior first-best solution exists.

5.2. Optimal Incentives for the Three Networks (Noncontractible Effort)

5.2.1. Self-Routing. When effort t is not contractible, the principal maximizes

$$V^S = \max_{0\leq p\leq 1,\, w,\, b} \frac{v-b-\bar{F}(t)(pc_I+(1-p)c_E)-C(p)}{t+p\bar{F}(t)r} - w,$$

$$\text{subject to } \frac{b}{t+p\bar{F}(t)r}+w-a\geq 0 \quad \text{(IR)},$$

$$t\in \arg\max_{t'\geq \underline{t}} \left\{\frac{b}{t'+p\bar{F}(t')r}+w-a\right\} \quad \text{(IC)}.$$

It turns out that the agents have the same optimal response as in the limited demand 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:

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

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

Doing this 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, it is independent of whether the agents are continuously busy or have idle time. The following lemma states the result.

Lemma 6. *Under unlimited demand and self-routing, the agents' unique optimal effort is t^S as defined 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

$$V^D = \max_{0\le p\le 1,\, w,\, b} \frac{1}{2t_1}[v - b - \bar{F}(t_1)(pc_I + (1-p)c_E) - C(p)] - \frac{w_1 + w_2}{2},$$

subject to

$$\frac{(1 - p\bar{F}(t_1))b}{t_1} + w_1 - a \ge 0 \quad \text{(IR1)},$$

$$\frac{p\bar{F}(t_1)}{t_1}(b - ar) + w_2 \ge 0 \quad \text{(IR2)},$$

$$t_1 \in \arg\max_{t' \ge \underline{t}} \left\{ \frac{(1 - p\bar{F}(t'))b}{t'} + w_1 - a \right\} \quad \text{(IC1)}.$$

To analyze the optimal effort, it is useful to define

$$\bar{p} = \frac{1}{\underline{t} f(\underline{t}) + 1}.$$

Lemma 7. *Under unlimited demand and dedicated routing, the agents' unique optimal effort is*

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

Different from the limited demand case, Agent 1's optimal effort does not depend on b. Because Agent 1 is continuously busy under unlimited demand, 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 effective throughput $(1 - p\bar{F}(t))/t$, which is not affected by b.

5.2.3. Cross-Routing. Unlike in the case of limited demand, we now need to consider only the queueing dynamics of the two rework queues (because the new job queues are nonempty with probability 1 in steady state). Recall the three stationary average fractions of time defined earlier: π_i^0 (being idle), π_i^n (working on new jobs), and π_i^r (working on rework). In steady state, π_i^0 equals zero and

$$\pi_i^n + \pi_i^r = 1,$$

$$\lambda_{ji}\pi_j^n = \mu^r \pi_i^r,$$

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

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

where $\rho_i \equiv \rho_i(t_i) \equiv p\bar{F}(t_i)(r/t_i)$. Agent i's long-run average utility rate is as follows:

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

Lemma 8. *Under unlimited demand and cross-routing, if $p > \bar{p}$, an interior symmetric Nash equilibrium (t^C, t^C) exists and is determined by*

$$p[t^C f(t^C) + \bar{F}(t^C)]\left[\rho(t^C)\left(1 - \frac{r}{t^C}\right) + 1 \right] + \rho(t^C)^2 - 1 = 0. \tag{6}$$

The existence condition is identical to the interior effort condition under dedicated routing. In addition, similar to dedicated routing, the equilibrium effort depends only on p. Because we focus on a symmetric Nash equilibrium, we safely suppress the subscripts distinguishing the agents. The principal's problem becomes

$$V^C = \max_{0\le p\le 1,\, w,\, b} \frac{v - b - \bar{F}(t)(pc_I + (1-p)c_E) - C(p)}{t + p\bar{F}(t)r} - w,$$

$$\text{subject to} \quad \frac{b}{t + p\bar{F}(t)r} + w - a \ge 0 \quad \text{(IR)},$$

$$t = t^C \quad \text{(IC)}.$$

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

In §4, we showed 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. Throughout this section, we implicitly assume that the existence condition in Lemma 8 is satisfied so that we have a meaningful definition of t^C to conduct the comparison.

Proposition 2 (Unlimited Demand). *The first-best solution* $\{p^{FB}, t^{FB}\}$ *cannot be achieved by any of the three rework routing schemes. Furthermore,* $t^S(p) < t^{FB}(p)$.

The conventional self-routing scheme induces lower effort than the first-best situation at any inspection precision p. As a result, self-routing cannot achieve the first best. Dedicated routing and cross-routing cannot attain the first best either. It is the lack of an additional incentive lever (i.e., piece rate b) that causes the first best not to be attainable—notice that the agents' effort is only affected by p in all three routing schemes. Next we compare the performance of the three routing schemes and summarize the results in the right column of Table 1.

Corollary 3 (Unlimited Demand). $t^C(p) > t^D(p) > t^S(p)$ *and* $Q^C(p) > Q^D(p) > Q^S(p)$.

Similar to the limited demand 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. Because rework is favorable, each agent would like the other agent 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.

5.3.1. 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 he also forces agent j to spend a larger fraction of his time on new jobs and thus generate more rework for agent i, keeping agent j's effort unchanged. Analytically,

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

$\partial\pi_j^n/\partial t_i > 0$ illustrates the workload externality imposed on agent j when agent i increases his first-pass effort. Because 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.

Corollary 4 (Unlimited Demand). *At any given wage rate,* $b^C(p) > b^S(p)$*; at zero wage rate,* $b^C(p) > b^D(p) > b^S(p)$.

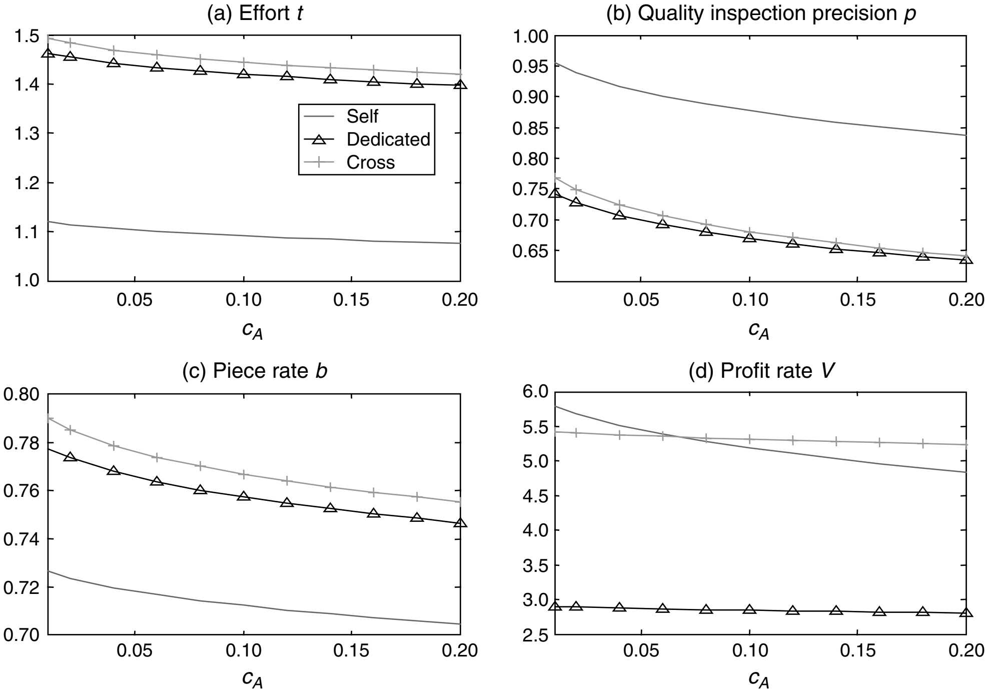
Lemma 4 states that the principal pays a higher piece rate to compensate for the higher effort that the agents exert under cross-routing and dedicated routing. More interestingly, using this piece-rate ranking, we can show that the effort equilibrium of cross-routing is a prisoner's dilemma.

5.3.2. 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 the agents a strictly positive utility rate because $b^C(p) > b^S(p)$ (when an interior effort equilibrium exists)—a better outcome for both agents than the equilibrium outcome that renders a zero utility rate for both agents. Because $f(t^S) = 1/pr$,

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

The last inequality follows from the fact that $r \le t^S$. Therefore, agent i has an incentive to unilaterally deviate from the cooperative outcome. This prisoner's

Figure 3 Under Unlimited Demand, Equilibrium Performance Depends on the Appraisal Cost: High Gross Margin ($v = 10$)



Notes. $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = c_A p^2/(1-p)$, $a = 0.5$, $r = 0.5$, $w = 0$.

dilemma induces higher first-pass effort and thus leads to a higher-quality output.

We caution that the superior incentive 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 et al. 1995), a collusive outcome (t^S, t^S) 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. We now compare the principal's profit rate.

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

(i) *If* c_I, c_E, *and* C'' *are sufficiently large,* $V^C > V^D > V^S$,

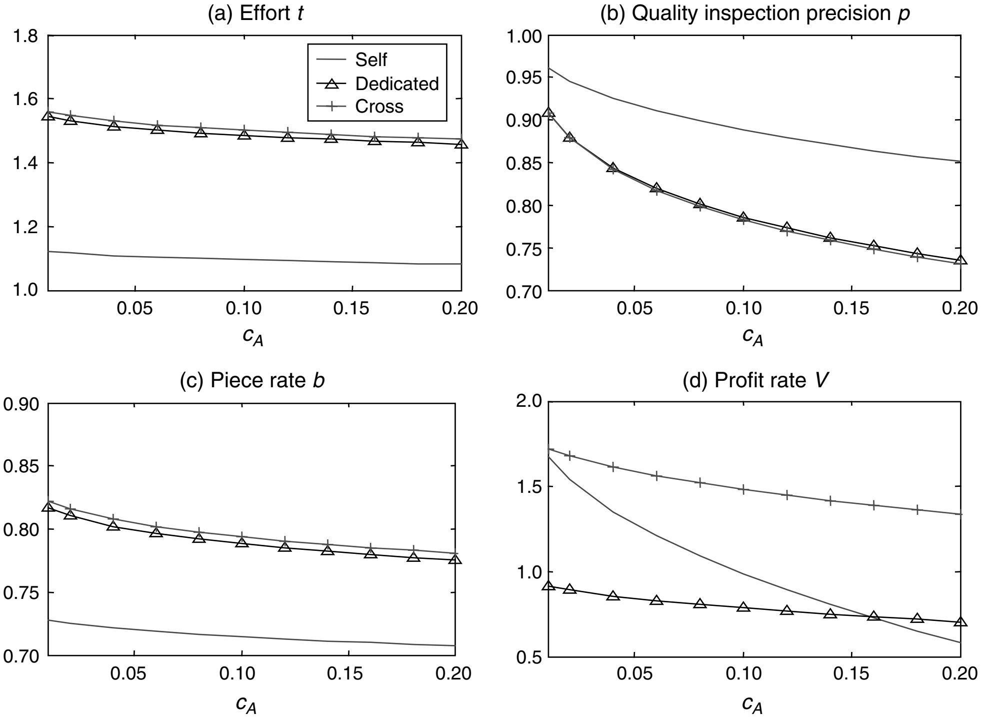
(ii) *If* v *is sufficiently large,* $V^S > \max\{V^D, V^C\}$.

In general, the ranking of the three schemes depends on the parameters, but when certain parameters are sufficiently large (the proof shows how to quantify this), the ranking is clear. The conditions in Proposition 3 suggest that under unlimited demand 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. Total expected service time (including rework time) is convex in first-pass effort, so increasing effort beyond t^S (i.e., the effort that minimizes the total expected service time) reduces effective throughput and consequently lowers the revenue rate. Therefore, cross-routing underperforms self-routing when the gross margin 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.[7]

We illustrate these effects by numerical examples. When the gross margin is high (Figure 3), there exists a threshold of c_A (c_A is a scale factor of $C(\cdot)$ and also a measure of C'') below which self-routing generates the highest profit rate. In contrast, when the gross margin

[7] It is possible that achieving the rank order $V^C > V^D > V^S$ requires high quality costs that lead to negative profit rates for all three schemes. Under these circumstances, self-routing dominates the other two schemes for the feasible range of quality costs that yields positive profit rates.

Figure 4 **Under Unlimited Demand, Equilibrium Performance Depends on the Appraisal Cost: Low Gross Margin** ($v = 4$)



Note. $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = c_A p^2/(1-p)$, $a = 0.5$, $r = 0.5$, $w = 0$.

is low (Figure 4), cross-routing always dominates the other two schemes. In addition, there exists a threshold of c_A above which $V^C > V^D > V^S$. These results are consistent with Proposition 3.

5.4. Equilibria with Many Agents

As was true for the case of limited demand, the system scales up proportionally under self-routing and dedicated routing, so their incentive effects remain unchanged. In contrast, we will show that the incentive effects of cross-routing decline as the system size increases. Under cross-routing, agent i's long-run average utility rate becomes

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

Lemma 9. *Under unlimited demand and cross-routing, if $p > \bar{p}$, an interior symmetric Nash equilibrium exists and the equilibrium effort t_N^C is determined by*

$$p[t_N^C f(t_N^C) + \bar{F}(t_N^C)]\left[\rho(t_N^C)\left(1 - \frac{1}{N-1}\frac{r}{t_N^C}\right) + 1\right] + \frac{1}{N-1}\rho(t_N^C)^2 - \frac{N-2}{N-1}\rho(t_N^C) - 1 = 0. \qquad (7)$$

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

Lemma 10. *For any $N \geq 2$, $t_N^C(p) > t^D(p)$, where $t^D(p)$ is the optimal effort under dedicated routing, as defined in Lemma 7.*

This lemma shows that the incentive effect of cross-routing remains 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 establish a monotonicity property that the incentive effect of cross-routing decreases in the size of the system. The condition involves *increasing generalized failure rate* (IGFR). Lariviere and 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 11 (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 strictly monotone decreasing in N.*

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). A commonly used distribution that satisfies the conditions is exponential: $F(t) = 1-\beta e^{-(t-\underline{t})}$. Because it has a 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.

Notice that if we take N to the limit, the equilibrium condition reduces to $f(t) = (1-p\bar{F}(t))/pt$, 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 epsilon equilibrium, which we adopt here.

Definition 1 (ε-Equilibrium). Let a real number $\varepsilon \geq 0$. An N-tuple $\hat{t} \in [\underline{t}, \infty)^N$ is an ε-equilibrium if for any agent i and any strategy $t_i \in [\underline{t}, \infty)$, $U_i(t_i, \hat{t}_{-i}) - U_i(\hat{t}_i, \hat{t}_{-i}) \leq \varepsilon$.

In other words, ε-equilibrium describes the strategy profile that is within ε of the best payoff of each agent. Notice that ε-equilibrium is a weakened notion of a Nash equilibrium. If $\varepsilon = 0$, the definition reduces to that of the Nash equilibrium. 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 that acceptance of an epsilon equilibrium reflects the bounded rationality of individual agents.

Proposition 4. *For any real number $\varepsilon > 0$, there is an integer N_ε such that, for all $N \geq N_\varepsilon$, there exists an ε-equilibrium in which each agent plays t^D as defined 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 for two agents. Earlier we showed 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 a large game with interchangeable agents. The strategic interactions thus diminish to none and the incentive effect of cross-routing reduces to implicit punishment—the incentive driver of dedicated routing. The diminishing incentive effect 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. Dependent Rework Time

We now relax the assumption that the rework time has a constant mean by letting it vary with first-pass effort. Consider an environment where job completion requires a certain amount of total effort; i.e., less first-pass effort leads to more rework effort. To be precise, let r depend on the first-pass effort t in a linear way, i.e., $r = \tau - t$, where τ is a positive constant. In this section, we will pass directly to the results without laying out the optimization problems. 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.

Under limited demand, we show that dedicated routing and cross-routing can attain first best while self-routing cannot. Under unlimited demand, although the agents' problems remain well behaved, the comparison of the three rework routing schemes becomes analytically less tractable. To test the robustness of our main results derived earlier, we conducted a numerical study. When the gross margin is high, Figure 5 shows that there exists a threshold of c_A below which self-routing leads to the highest profit rate. In contrast, when the gross margin is low, Figure 6 shows that cross-routing always dominates the

Table 2 **Agents' Optimal Effort Under Dependent Rework Time (Assuming Interior Solutions)**

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

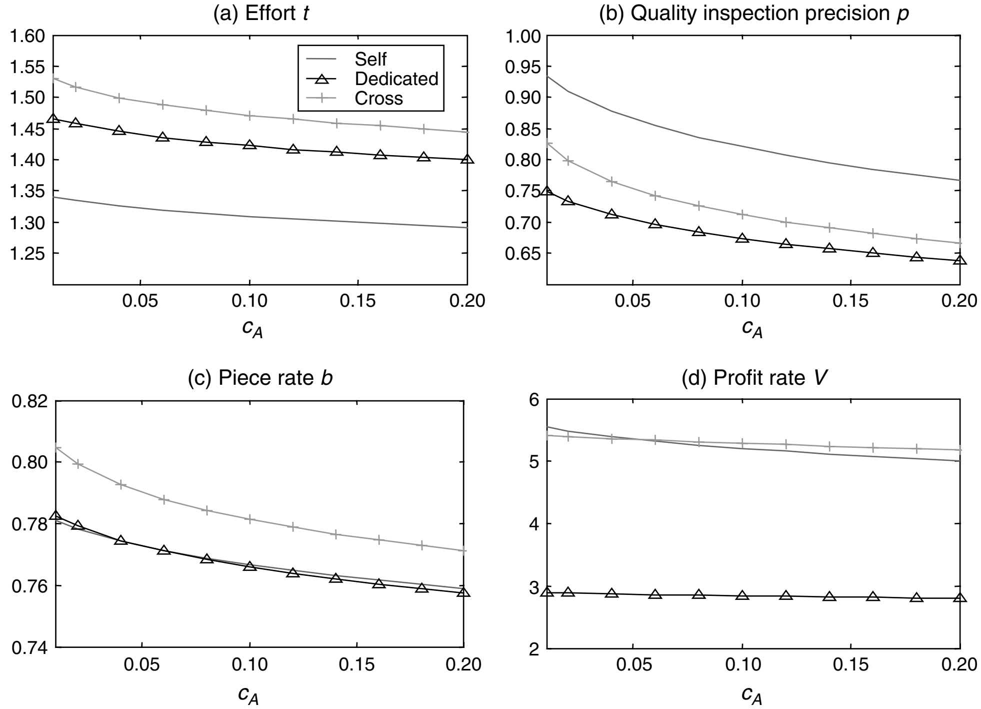
other two routing schemes: $V^C > V^D > V^S$. These numerical findings are consistent with the results in Proposition 3.

7. Conclusions and Discussion

This paper investigates how incentives and judicious rework routing can improve quality and the profitability of a firm using a principal-agent model integrated into a multiclass queueing network. We demonstrate that traditional self-routing of rework is suboptimal in 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 but also on demand levels, revenues, and quality costs. The novel cross-routing scheme is applicable in both a manufacturing and a service operations environment. The merit of this scheme lies in the fact 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 effect of cross-routing reduces monotonically and approaches that of dedicated routing under certain conditions.

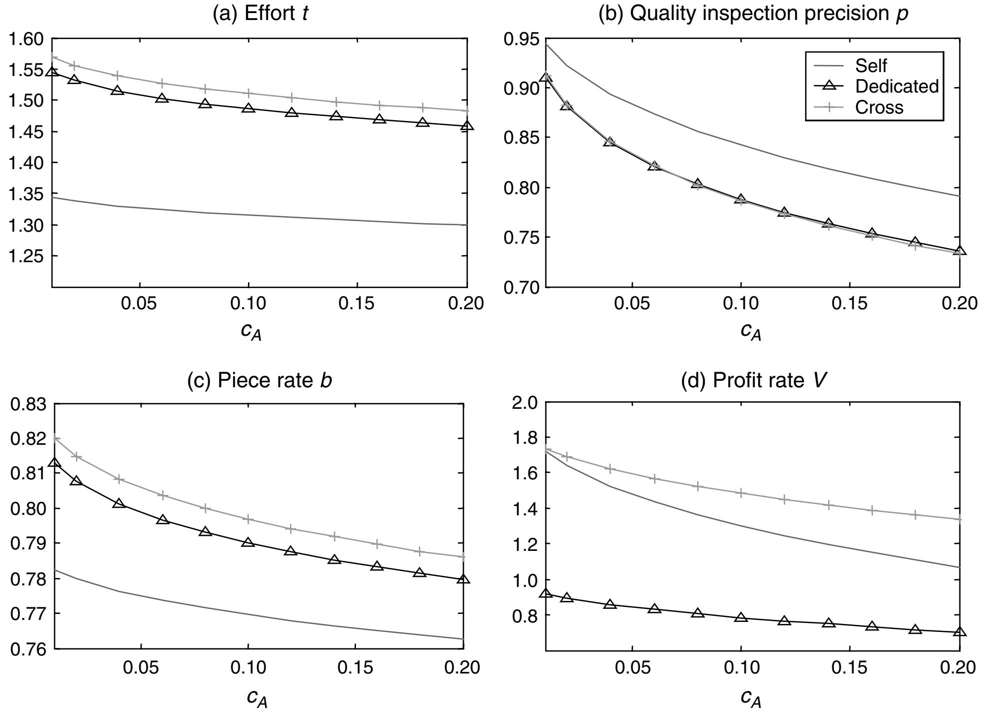
It is worth noting that agents being continuously busy (i.e., the unlimited demand regime) is a crucial condition that gives cross-routing additional incentive power compared with dedicated routing. Given that system throughput is exogenous under limited

Figure 5 **Under Unlimited Demand, Equilibrium Performance Depends on the Appraisal Cost: High Gross Margin** ($v=10$)



Notes. $F(t)=1-e^{-3(t-1)}$, $c_I=0.5$, $c_E=10$, $C(p)=c_A p^2/(1-p)$, $\tau=2$, $a=0.5$, $w=0$ (assuming dependent rework time).

Figure 6 Under Unlimited Demand, Equilibrium Performance Depends on the Appraisal Cost: Low Gross Margin ($v = 4$)



Notes. $F(t) = 1 - e^{-3(t-1)}$, $c_I = 0.5$, $c_E = 10$, $C(p) = c_A p^2/(1-p)$, $\tau = 2$, $a = 0.5$, $w = 0$ (assuming dependent rework time).

demand yet endogenous under unlimited demand, it is not surprising that the asymptotic interpretation of the limited demand regime is not equivalent to that of the unlimited demand regime. Addressing the discontinuity between our two model regimes would require adding another layer of complexity by making demand dependent on the service rates and qualities implemented by the principal and agents.

However, let us discuss why we believe our mode of analysis is sufficient for our objectives. Although a fully utilized operation with unlimited demand may seem unrealistic, its analysis highlights key causal relationships and offers insights applicable to stable systems that are heavily loaded. Indeed, recall that we have focused on long-run average effects in this paper for reasons of tractability. But busy periods in heavily loaded systems are sufficiently long for the system dynamics and short-run profits and incentive effects to resemble those under unlimited demand. The workload externality effect generated by the agents' strategic effort decisions becomes nonnegligible for a principal who cares about short-run system performance.

We have made two methodological contributions to the agency and operations management literature. First, we study a multiagent principal-agent model in a multiclass queueing network with endogenous queues (recall that 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 queues with endogenous arrival rate in a principal-agent framework. Our application of the epsilon equilibrium concept to study an operational system with many agents is also a novel approach. Second, we embed the quality-quantity trade-off in a single-dimensional decision variable, i.e., the average processing time per job. In contrast to the multitask principal-agent model, our approach is applicable 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, because of the inherent variability in queueing networks, risk aversion cannot easily be incorporated, given that we conduct a 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 or 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 in a straightforward way. We believe this will change our result in one direction: the principal would be more reluctant to induce quality effort because higher first-pass effort may lead 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 workers to learn from their own mistakes. Because the workers are independent, self-routing may also be conducive for teamwork: the workers can share their knowledge and experiences on quality improvement without jeopardizing their own performance and compensation. Because the utilization of agents is not balanced in dedicated routing, cross-routing may be preferred even though the two routing schemes may have the same incentive effects. However, cross-routing loses the specialization benefit of dedicated routing. Moreover, because of the "competition" between workers, cross-routing may dampen their interest in teamwork such as sharing quality improvement ideas with coworkers.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

The authors thank Kevin Wilson of Memphis Auto Auction for bringing cross-routing to their attention. Three anonymous referees and the editors provided suggestions that significantly improved the paper. The authors also greatly benefited from suggestions by Gad Allon, Sandeep Baliga, James Dana, Harry Groenevelt, Michael Harrison, Wallace Hopp, Seyed Iravani, Johannas Horner, Ehud Kalai, and Martin Lariviere.

References

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

Benjaafar, S., E. Elahi, K. L. Donohue. 2007. Outsourcing via service competition. *Management Sci.* **53**(2) 241–259.

Cachon, G., P. T. Harker. 2002. Competition and outsourcing with scale economies. *Management Sci.* **48**(10) 1314–1333.

Cachon, G. P., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Sci.* **53**(3) 408–420.

Conway, R., W. Maxwell, J. O. McClain, L. J. Thomas. 1988. The role of work-in-process inventory in serial production lines. *Oper. Res.* **36**(2) 229–241.

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

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

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

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

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

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

Hasija, S., E. J. Pinker, R. A. Shumsky. 2008. Call center outsourcing contracts under information asymmetry. *Management Sci.* **54**(4) 793–807.

Holmstrom, B., P. Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *J. Law, Econom., Organ.* **7** 24–52.

Juran, J. M., F. M. Gryna. 1993. *Quality Planning and Analysis: From Product Development Through Use*. McGraw-Hill, New York.

Kekre, S., U. S. Rao, J. M. Swaminathan, J. Zhang. 2003. Reconfiguring a remanufacturing line at Visteon, Mexico. *Interfaces* **33**(6) 30–43.

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

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

Lazear, E. P. 2000. Performance and productivity. *Amer. Econom. Rev.* **90**(5) 1346–1361.

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

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38**(5) 870–883.

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

Parlakturk, A. K., S. Kumar. 2004. Self-interested routing in queueing networks. *Management Sci.* **50**(7) 949–966.

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

Radner, R. 1979. Collusive behavior in noncooperative epsilon-equilibria of oligopolies with long but finite lives. *J. Econom. Theory* **22** 136–154.

Ren, Z. J., Y.-P. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Sci.* **54**(2) 369–383.

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

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

Spearman, M. L., M. A. Zazanis. 1992. Push and pull production systems: Issues and comparisons. *Oper. Res.* **40**(3) 521–532.

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

Van Oyen, M. P., E. G. S. Gel, W. J. Hopp. 2001. Performance opportunity for workforce agility in collaborative and noncollaborative work systems. *IIE Trans.* **33** 761–777.