

Informativeness, Incentive Compensation and the Choice of Inventory Buffer*

Stanley Baiman^a

Serguei Netessine^b

Richard Saouma^c

October 2009

*This paper has benefited from the comments of workshop participants at the University of Wisconsin – Madison, the University of Iowa, the University of Houston 2007 Accounting Symposium, the 5th Accounting Research Workshop (Fribourg, Switzerland), the University of Chicago, the University of Pennsylvania, Tel Aviv University, Hebrew University (Jerusalem), Norwegian School of Economics and Business, and the Catholic University of Portugal, especially those of R. Balakrishnan, P. Berger, R. Cazier, D. DeJong, J. Gerakos, F. Gjesdal, J. Glover, T. Hemmer, E.M. Matsumura, M. Penno, S. Radhakrishnan, K. Sivaramakrishnan, and J. Stecher. We are particularly indebted to Jack Hughes, Cathy Schrand, the Editor and two anonymous referees for many helpful suggestions and to Romanos Malikiosis for his research assistance.

^aDepartment of Accounting, The Wharton School, The University of Pennsylvania, Corresponding Author

^bDepartment of Operations and Information Management, The Wharton School, The University of Pennsylvania

^cDepartment of Accounting, The Anderson School of Management, UCLA

Abstract

Previous research in Management Accounting and Economics has noted the potential for complementarities between the firm's performance measurement system and its other organizational design choices. We add to this literature by studying how the informativeness and incentive properties of a performance metric can be influenced by one particular organizational design choice – the size of the firm's inventory buffers. We model a manufacturing setting in which an agent manages a workstation that processes intermediate units. As intermediate units arrive, they are stored in an inventory buffer until the agent can process them. The buffer can hold a maximum number of intermediate units – its *buffer size*. If an intermediate unit arrives while the buffer is full, the unit is diverted and the agent loses the opportunity to process it. If the workstation is ready to process another unit but the buffer is empty, then it must remain idle until the next intermediate unit enters the buffer. The agent is compensated on the basis of his workstation's throughput. We demonstrate how the choice of inventory buffer size can affect the informativeness of the performance metric and the incentive compensation wage necessary to motivate the agent. Further, we characterize the conditions under which reducing the inventory buffer enhances/degrades the informativeness of the performance metric, that is, mitigates/exacerbates the agent's incentive problem.

I. Introduction

Much of the agency theory work in Managerial Accounting has studied the relation between information system properties (e.g., informativeness, precision, congruity, etc.) and optimal incentive compensation. This work typically treats these information system properties as primitives or dials on a control panel that are directly manipulated, subject perhaps to an out-of-pocket cost of doing so. Yet in practice, these information system properties may also be indirectly affected by the firm's other organizational design choices, such as: job assignments, product design, and production technology. As (Hemmer 1998, pp. 321-322), states, "...the value of a performance measure is determined not simply by its congruity and precision but by its influence on the optimal organizational design....Much of the recent theoretical accounting literature...has largely ignored complementarities between performance measures and organizational design."

In this study, we address Hemmer's observation and study how the informativeness and incentive properties of a performance metric can be influenced by one of the firm's organizational design choices – the size of its inventory buffers. We model a manufacturing setting in which an agent manages a workstation that processes intermediate units. As intermediate units arrive, they are stored in an inventory buffer until the agent can process them. The buffer can hold a maximum number of intermediate units – its *buffer size*. The buffer size represents the workstation's maximum order backlog. If an intermediate unit arrives while the buffer is full, the unit is diverted and the agent loses the opportunity to process it. If the workstation is ready to process another unit but the buffer is empty, then it remains idle until the next intermediate unit enters the

buffer. While we frame the analysis in terms of a manufacturing setting, the analysis is also applicable to service operations such as call centers which can have a maximum number of calls awaiting service.

We assume that the agent is compensated on the basis of his throughput; i.e., the *output per unit of time* of his workstation. Consistent with Hemmer's observation, we show that the principal's choice of buffer size indirectly affects the informativeness of, and consequently the optimal compensation weight on, throughput. The buffer size affects both the probability that the agent's buffer will be full when an intermediate unit arrives, causing the agent to lose the opportunity to complete the unit (this is referred to as *blocking*), and the probability that the buffer will be empty when the agent is free, causing the agent to forego further production until another intermediate unit arrives (this is referred to as *starving*). Both probabilities affect the marginal productivity of the agent's effort with respect to throughput (his performance metric) and hence, the extent to which that effort is reflected in throughput. Further, we not only demonstrate that the principal's choice of buffer size affects the informativeness of throughput with respect to the agent's effort, but that the effect on informativeness can be non-monotonic. Much of the Accounting and Operations Management literature has emphasized two benefits of reducing inventory buffer size: the reduction of inventory holding costs and the increase in ability to uncover mistakes and determine when the underlying exogenous production process is "out of control". Our non-monotonicity result indicates that there can be an offsetting incentive cost to reducing inventory buffer size.

Next we discuss the relevant literature on organizational design and incentives. In Section II we introduce the model and in Section III we present the results. Section IV discusses extensions and concludes.

A number of articles in both the Accounting and Economics literatures have considered the effect of organizational design choice on the design of optimal incentives. Among the issues considered are: job design ((Holmstrom and Milgrom 1991), (Balakrishnan et al. 1998) and (Riordan and Sappington 1987)); the choice of production systems and technologies ((Milgrom and Roberts 1995), (Hemmer 1998) and (Hemmer 1995)); the hierarchical structuring of work ((Melumad et al. 1995)); routing schemes for product rework ((Lu et al. 2006)); and production bottlenecks ((Datar and Rajan 1995) and (Gietzmann and Hemmer 2002)).¹

Similar to our paper, (Hemmer 1998), (Hemmer 1998), and (Gietzmann and Hemmer 2002) examine how different workflow arrangements between agents affect the information available for contracting, and the incentives facing agents. Our work is distinct from these papers, which do not consider buffer size as a choice variable of the principal. (Nagar et al. 2009) does examine the role of inventory buffers in agency problems, although in their model, unlike ours, buffers are filled by agents to signal private information and buffer size is not a choice variable. In contrast, Alles et al. (1995), like the present work, focuses on the effect that the choice of buffer size can have on the informativeness of performance metrics. The major difference between our work and theirs is that they do not formally model that effect, but rather assume a monotonic relation. In contrast, we formally model the inventory process and derive a non-

¹ Also somewhat related to the present work is (Cremer 1995) which examines the incentive effects on both the buyer and supplier when the buyer goes to a zero-inventory system for the supplier's product. However, the two motivations are different in that Cremer's is based on a double moral hazard model.

monotonic relation between buffer size and the informativeness of throughput. However, for reasons of tractability, unlike Alles et al., we assume that the agent is risk-neutral.

II. The Model and Initial Analysis

The firm consists of a risk-neutral principal and agent who, at the start of the period, agree to a contract. The agent is hired to set up a workstation that processes incoming, intermediate units (e.g. testing completed computers or welding automotive chassis). The intermediate units arrive stochastically, with a known mean arrival rate of λ , at the workstation's incoming inventory buffer, whose capacity is $b \in \mathbb{N}$, allowing a maximum of $b - 1$ units to be held in front of the workstation while the workstation processes 1 unit. As described earlier, if an intermediate unit arrives and the buffer is not full, the unit is added to the inventory. However, if the buffer is full, blocking occurs and intermediate units cease arriving until there is space in the buffer, whereupon the intermediate units begin arriving at the same stochastic rate as before.²

Once the contract is agreed to, but before the workstation begins processing, the agent chooses his personally costly set-up effort r , which affects the rate at which his workstation can process the intermediate units. That effort is unobserved by the principal and is subject to moral hazard. The arriving intermediate units are sufficiently heterogeneous that the agent's effort does not perfectly control the rate at which the workstation processes the incoming intermediate units. Rather, the intermediate units are processed at a stochastic rate with mean $r(1 + h)$, where $h \in (-1, \infty)$ represents the

² Alternatively, one can interpret the model as one in which the intermediate units continue arriving while the inventory is blocked but are diverted elsewhere. What is important is that in both cases, blocking imposes an opportunity cost on the agent because he loses the chance to process the intermediate units that might have been added to the buffer had it not been full.

productivity-enhancing resources the principal has allocated to the agent. Examples of the latter include: machine and tool upgrades, improved working conditions, training, etc. As the workstation finishes processing an intermediate unit, the agent reaches into the buffer for the next unit on which to work. If the buffer is empty, starving occurs and the workstation is idle until the next intermediate unit arrives. See Figure 1 for a graphical description of the process.

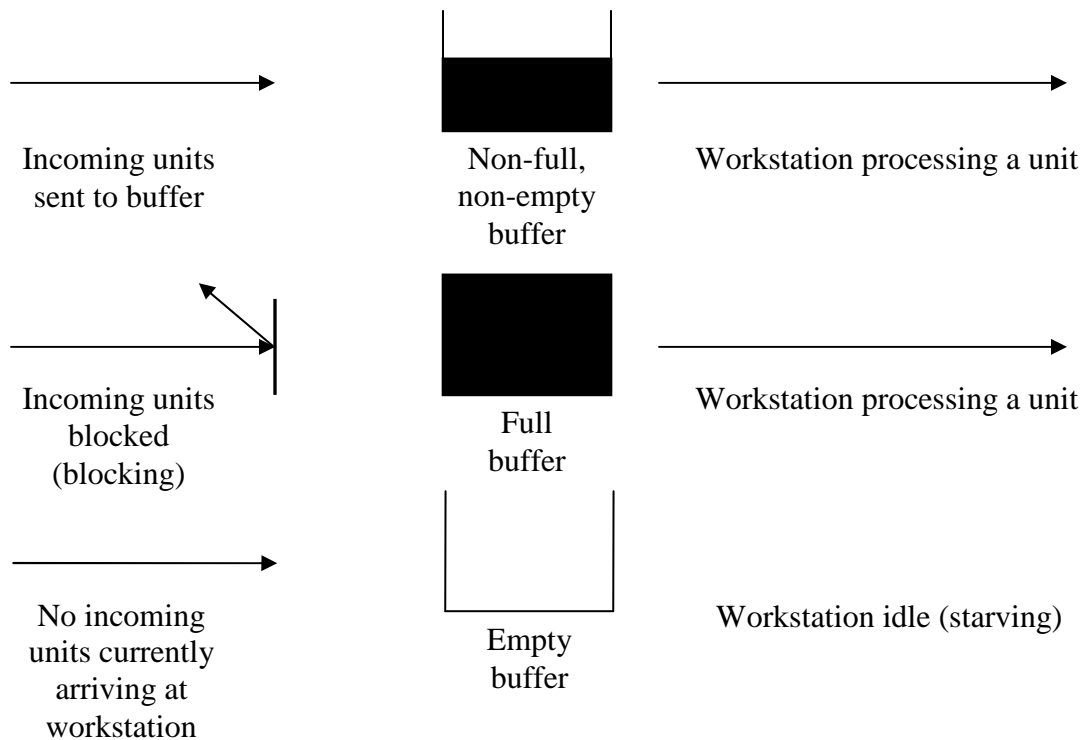


Figure 1: In the top panel, the buffer is not full, incoming units are added to the buffer, and the workstation starts processing a unit from the buffer once it completes the current unit. In the middle panel, the buffer is full, incoming units are blocked from entering the buffer until the workstation finishes processing its current unit, and the agent withdraws the next intermediate unit from the buffer (blocking). In the bottom-most panel, the buffer is empty and hence the workstation is idle (starved) until the next unit arrives.

The principal chooses both the agent's compensation scheme and the buffer size, b . We restrict our attention to linear compensation schedules for the agent consisting of a fixed payment, w_0 , plus an incentive wage, w , based on realized throughput – the number of units processed per unit of time.^{3 4}

Lastly, we assume that the marginal cost associated with each unit of buffer capacity is $c > 0$, the agent's marginal cost of effort is $a > 0$, and we normalize the marginal cost of h at 1 dollar per unit of resource.

All notation and assumptions are as follows:

λ = the mean arrival rate of the intermediate units. The arrival rate is Poisson distributed, implying that the time between arrivals is distributed according to an exponential distribution with mean of λ^{-1} .⁵

r = the level of effort the agent exerts at private cost, ar .

h = the amount of productivity-enhancing resources allocated to the agent. The principal's out-of-pocket cost is h .

$r(1 + h)$ = the mean rate at which the agent's workstation processes the intermediate units. The actual processing time is distributed according to an exponential distribution.

³ Given that the basic properties of queuing systems are stated in terms of rates (i.e., arrivals per unit of time, number of units processed per unit of time) we assume that the principal compensates the agent on throughput per unit of time rather than total throughput. Linear compensation schemes are frequently used in practice (see (Berg and Fast 1975) and (Hall et al. 2000)) and are used extensively in the analytical literature; e.g., (Feltham and Xie 1994), Alles et al. (1995), Hemmer (1995), Hemmer (1998), and (Gietzmann and Hemmer 2002).

⁴ In the Conclusion, we discuss the possibility of contracting on the inventory level.

⁵ In an exponential distribution, the same parameter that represents the mean also represents the variance. For simplicity we refer to the choice of this parameter as the choice of the mean.

b = the maximum number of units the station can store (buffer size). The principal incurs the capacity cost cb each period.

K = the steady-state mean throughput per period, with \bar{K} denoting the realized throughput.

w, w_0 = the incentive payment made to the agent per unit completed per period, and the fixed salary payment respectively.

Notice that in our model there are two sources of noise that make it difficult for the principal to infer the agent's effort from throughput. As is common in agency problems, the relation between the agent's effort and the time required for the workstation to process an intermediate unit is stochastic. In addition, the rate at which units become available to be processed is also stochastic. Thus, even if the relation between the agent's effort and the time required to process an intermediate unit were deterministic, instantaneous throughput would still be stochastic in the agent's effort. Placing an inventory buffer in front of the agent's workstation is one way of dampening out the effect of the variation in the arrivals; however, as we will show, such dampening can also reduce the informativeness of throughput with respect to the agent's effort.

Our assumptions regarding arrival and processing rates follow the standard $M/M/1/b$ queuing model with finite buffers used in the Operations Management literature.⁶ The *transient* behavior of queuing systems with finite buffers cannot be described analytically, although the *steady-state* behavior of such systems can be

⁶ (Balsamo et al. 2001), (Berkley 1992), and (Groenevelt 1993), among others, cite the $M/M/1/b$ model as the most widely accepted modeling approach for production systems with finite buffers.

described in closed form.⁷ As a result, the Operations Management literature typically uses simulations to analyze the transient behavior of queuing models with finite buffers and closed-form analysis to examine steady-state behavior. In contrast, agency models have emphasized rewarding transient behavior, but by simplifying the production process (e.g., using the LEN model) to achieve closed-form solutions.⁸ In order to incorporate stochastic intermediate unit arrival rates, stochastic processing rates, and finite inventory buffers in our analysis, we must accept a number of simplifying assumptions, which we discuss next.

First, we focus exclusively on the problem's *steady-state*. That is, we assume that the principal is interested in maximizing her *steady-state* expected profit. In turn, we assume that the agent selects his effort to maximize his *steady-state* expected utility.⁹ Second, focusing on the steady-state implies that we must also assume that the agent cannot intervene while the process is operating; for example, he cannot vary his effort in response to changes in the number of intermediate units in the buffer, because such variation would prevent the system from attaining a steady-state equilibrium. This motivates our earlier assumption that the agent's effort is primarily involved in setting up the workstation before production begins this period.¹⁰ Third, even with these simplifications, for reasons of tractability, we need to assume risk-neutrality on the part of the agent. However, we do assume that the agent is subject to a limited liability

⁷ The process attains steady-state when the probability distributions over inventory levels and throughput are no longer functions of the starting conditions, and no longer vary over time. That is, while the realized inventory level and throughput at any moment are still uncertain, the probability distributions which describe them do not vary over time.

⁸ For exceptions see (Balachandran and Radhakrishnan 1996), (Radhakrishnan and Balachandran 1995) and (Radhakrishnan and Balachandran 2004), who use queuing models similar to ours to analyze the role of cost allocation in resolving congestion problems.

⁹ (Lu et al. 2006) and (Kim et al. 2007) restrict their analysis to steady-state for the same reason.

¹⁰ An example would be the agent who sets up and programs the workstation that welds chassis in the automotive production line. As long as the workstation is in control, the agent does not intervene.

constraint. In particular, we assume that the principal must leave the agent with a minimum level of utility regardless of realized output.¹¹

As noted, the workstation's realized output per unit of time or throughput, \tilde{K} , is stochastic, though the *mean* steady-state throughput, K , can be represented as (Hopp and Spearman 2001):

$$K(b, r) = \lambda(1 - p_b) = \lambda \frac{1 - \left(\frac{r(1+h)}{\lambda} \right)^{-b}}{1 - \left(\frac{r(1+h)}{\lambda} \right)^{-(b+1)}} \quad (1)$$

where p_b is the steady-state probability of the buffer being full; i.e., the steady-state probability of blocking. The workstation's steady-state mean throughput is thus equal to the unblocked mean arrival rate of intermediate units, (λ) , times the steady-state probability that the buffer can accept an additional unit when it arrives $(1 - p_b)$. In equilibrium, the workstation's average output rate is thus equal to the expected *effective* arrival rate of the intermediate units. The steady-state mean throughput is increasing in: r , the agent's effort; h , the level of resources allocated to the agent; b , the size of the input buffer; and λ , the mean arrival rate of the intermediate units. Equation (1) demonstrates that the agent's effort affects throughput via its effect on the probability of blocking, or equivalently, the probability of starving.¹²

¹¹ (Demougine and Garvie 1991) analyzes a model with a similar assumption. Alternatively, we could assume that, because of the agent's limited liability, the principal cannot pay either a negative piece rate or a negative fixed wage. Our results remain unchanged with this alternative formulation. The only effect is on the optimal fixed wage.

¹² The two probabilities are directly related in that the probability of starving is $p_s = 1 - \frac{\lambda}{r(1+h)}(1 - p_b)$, allowing (1) to be rewritten as $K(b, r) = r(1+h)(1 - p_s)$. Note that a change in the buffer affects the probability of starving in the same direction (i.e., increase/decrease) as its effect on the probability of

In designing the firm, the principal maximizes her expected steady-state profit with respect to her choice of (w, w_0, b, r) subject to satisfying the agent's Individual Rationality constraint and Incentive Compatibility constraints, where the latter ensures that the agent finds it in his best interest to choose the principal's desired effort level, r . This full profit-maximization problem can be solved in two stages. First is the cost minimization stage in which the principal finds the least costly (w, w_0, b) that will induce the agent to choose each feasible effort level r . This generates a feasible set $\{(w, w_0, b, r)\}$. Next is the profit-maximization stage, in which the principal chooses from this feasible set the (w, w_0, b, r) that maximizes her steady-state expected profit. The advantage of this two-stage approach is that analyzing the first stage is easier than analyzing the full profit-maximization problem, yet as noted in Grossman and Hart, 1983, one can still derive the properties of the optimal solution from the results of the first-stage analysis. In particular, any incentive effects of inventory buffer size found in the first-stage analysis will influence the principal's full profit-maximization problem.¹³ Consequently, our subsequent analysis focuses on analyzing the first step – minimizing the cost of inducing the principal's desired effort r^* .¹⁴

The first stage of the analysis can be written as:

blocking.

¹³ In Appendix A we provide several numerical examples to illustrate this point.

¹⁴ Other papers which use the Grossman-Hart approach to study situations in which the agent's obedient action is exogenously given include: (Gigler and Hemmer 2002), (Dutta and Gigler 2002) and (Arya and Glover 2008).

$$\begin{aligned}
 & \text{Min}_{\{w,b\}} wK(b, h, r^*) + w_0 + cb + h \\
 \text{(Program 1)} \quad & \text{s.t.} \\
 & w\tilde{K} + w_0 - ar \geq 0 \quad \forall \tilde{K} \quad (LL) \\
 & r^* \in \arg \max_r (wK(b, r) + w_0 - ar). \quad (IC)
 \end{aligned}$$

Constraint (LL) assures that the agent attains a non-negative utility for *any* realized throughput rate, \tilde{K} . Constraint (IC) ensures that the agent will select the principal's desired level of effort r^* , when he maximizes his steady-state expected utility. In the First-Best case, in which the agent's effort is not subject to moral hazard, the (IC) constraint can be ignored and, given that the minimum possible realized throughput is $\tilde{K} = 0$, constraint (LL) implies that $w_0 = ar^*$ and $w = 0$. That is, the principal pays the agent $w = 0$ and $w_0 = ar$ if and only if the obedient effort $r = r^*$ is observed. Hence the size of the buffer b plays no role in the agent's compensation in the First-Best solution. In the Second-Best setting, Constraint (LL) will again bind, implying that $w_0 = ar^*$. However, the (IC) constraint now requires that $w > 0$. Thus, w represents the agency cost per unit of throughput, and it is the behavior of w in b that captures the incentive effect of b in the Second-Best setting. Finally, note that while the principal could infer the agent's effort from the *mean* steady-state throughput rate, only the *realized* throughput rate, \tilde{K} , is observed, upon which the principal cannot deterministically derive the agent's choice of effort.

In the full profit-maximization problem in which the principal searches over $\{(w, w_0, b, r)\}$, she balances three effects in choosing b . Both the capacity cost of inventory (cb) and the expected revenue (via the mean throughput $K(b, r)$) are

increasing in b . These are the direct effects of the principal's choice of b . The principal's choice of b also has an indirect incentive effect. Constraint (IC) indicates that the choice of b influences the incentive compensation wage, w , required to induce the agent to choose each feasible action, r , and thus, which action the principal will choose to induce the agent to take in the second-stage analysis. The following Lemma characterizes this incentive effect and guarantees the validity of the First-Order Approach to the agent's choice:

Lemma 1: The agent's problem is concave in r . Further, the optimal piece-rate is:

$$w(b, r^*) = a \left(\frac{\partial K(b, r)}{\partial r} \right)^{-1} \Big|_{r=r^*}$$

which is non-negative and increasing in the principal's choice of r^* .

The optimal incentive wage is given by the agent's marginal cost of effort, divided by the sensitivity of mean steady-state throughput to his effort. The more sensitive the performance metric is to the agent's effort (i.e., the greater the agent's marginal productivity), the lower the piece-rate required to induce the agent to implement the obedient effort and the lower the agency cost of doing so. The basic idea is that the more sensitive K is to variations in r , the more likely it is that the principal can detect any deviation from the desired r^* ; therefore, the lower the incentive wage necessary to dissuade the agent from shirking. Henceforth, we refer to the sensitivity of throughput to the agent's effort, $\frac{\partial K(b, r)}{\partial r}$, as the *informativeness* of K with respect to the agent's

effort.¹⁵ As illustrated by (1), the agent's ability to affect K with his choice of effort is entirely driven by the agent's ability to change the probability of blocking and starving via his effort. Given that the wage required to induce the agent to take the desired action is inversely related to the informativeness of throughput, we turn our attention to characterizing the relation between the principal's choice of the buffer size b and the informativeness of throughput with respect to the agent's action.

III. Results

We will sometimes refer to the mean rate at which the workstation processes the intermediate units relative to their mean arrival rate, $r(1+h)/\lambda$, as ρ , the workstation's production schedule. We parameterize our discussion with respect to the production schedule, ρ , since we treat both the amount of help resources, h , and the mean arrival rate of the intermediate units, λ , as fixed.

Recall that in the agency model, the principal compensates the agent as if she were inferring the agent's effort from the performance metric (Holmstrom 1979). Consider first the case in which $\rho > 1$, implying that, on average, the workstation processes intermediate units more quickly than they arrive. If the buffer size, b , is sufficiently small, the resulting probabilities of blocking and starving are large, causing

¹⁵ Notice that our interpretation of informativeness and its relation to the agent's incentive compensation weight are somewhat different from the signal to noise ratio interpretation and relation in (Banker and Datar 1989). The reason for these differences are: 1) our agent is risk-neutral and, thus, the effect of noise on risk is irrelevant; and 2) with a (Grossman and Hart 1983) type analysis, the principal's desired effort for the agent does not vary with the informativeness of output. Our notion of informativeness and its relation to limited liability rents is most closely related to that in Laux (2001) and Laffont and Martimort (2002) p156.

throughput to be small. In this case, because of the small throughput, there are *few units per period of time available to the principal (i.e., a small sample size) on which to infer the agent's action choice*: therefore, throughput is relatively uninformative about r . By *increasing* the buffer, b , the principal can increase the number of units produced per unit of time available to the principal (i.e., throughput) on which to infer the agent's action. The increased number of intermediate units available for final processing from increasing b makes the mean throughput, K , more sensitive to, and hence informative of, the agent's effort choice. Therefore, when b is small, the principal can improve the informativeness of throughput by increasing b .

The above intuition is based on a small increase in b starting at a buffer which, given $\rho > 1$ (i.e., $r > \lambda$), is sufficiently small to severely restrict throughput because of blocking and starving. However, the effect of increasing the buffer on the informativeness of throughput differs once the buffer is already large enough so that blocking and starving are not severe. Increasing buffer size beyond some point must result in a *decrease* in the informativeness of throughput with respect to r . To see this, note that in the extreme, when b is infinite, there is zero probability of blocking and, given $\rho > 1$, the mean steady-state throughput is $K = \lambda$, which is independent of the agent's effort in excess of λ . Thus, when there is no blocking, K is *totally* insensitive to and uninformative about the agent's effort beyond the cut-off level λ . This implies that some probability of blocking is necessary for the throughput to be informative with respect to the agent's effort.

Recall that throughput can be informative about its two sources of variation: the rate at which the intermediate units arrive and the rate at which available units are processed by the workstation, whereas only the latter is determined by and informative about r . Increasing b beyond some point results in the relative amount of information about this second source of variation (on which the principal wants to base the agent's compensation) to decrease drastically. So when b is relatively large, the principal can increase the informativeness of throughput by *decreasing* b . This intuition is formalized in the following proposition:¹⁶

Proposition 1: For any fixed $\rho > 1$, the informativeness of throughput with respect to the agent's choice of effort, r , is either decreasing in the size of the buffer, $b \in \mathbb{R}^+$, or initially increasing in b and then decreasing in b .

Figure 2 illustrates Proposition 1 and the preceding intuition; when the buffer is relatively small (large), the informativeness of throughput is increasing (decreasing) in buffer size. The proposition also adds the possibility that the informativeness of throughput is monotonically decreasing in buffer size. This case arises because, no matter how small the buffer size ($b \geq 1$), there always exists a sufficiently large value of ρ that causes the probability of blocking and starving to be very small, causing the informativeness of throughput to decrease in buffer size. Using Lemma 1, Proposition 1 leads to the empirically testable result that the incentive wage rate (i.e., the marginal cost of effort divided by the informativeness of throughput) should be either increasing in buffer size or U-shaped.

¹⁶ All proofs are in Appendix B.

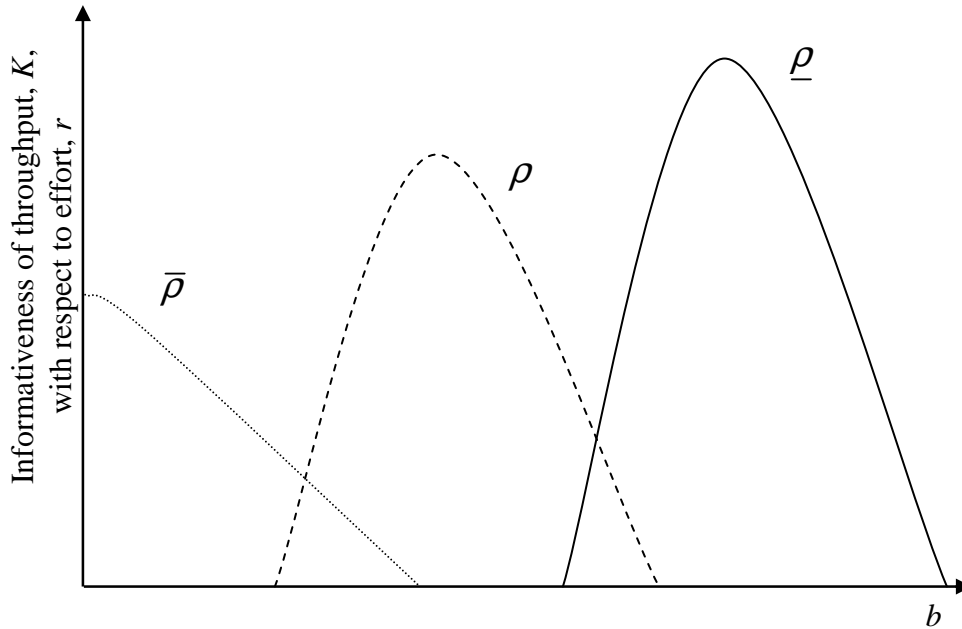


Figure 2: The level of informativeness is plotted for three different production schedules: ρ , with $\underline{\rho} < \rho < \bar{\rho}$. Note that additional buffers may always decrease the informativeness of throughput when the production schedule requires the agent's station to process incoming units sufficiently quickly, as is the case above with $\bar{\rho}$.

The intuition for Proposition 1 suggests that, when a relatively large buffer causes the probability of blocking to be very small, the principal would want to decrease the buffer, while when a relatively small buffer causes the probability of blocking to be very large, the principal would want to increase the buffer. Since the probability of blocking is always decreasing in ρ , intuition suggests that the optimal buffer size should also be decreasing in ρ . This logic is formalized in the following proposition:

Proposition 2: For $\rho > 1$, the buffer size, b , which maximizes the informativeness of throughput with respect to the agent's choice of effort, is decreasing in the production schedule, ρ .

To this point we have addressed only the case in which $\rho > 1$. When $\rho < 1$, on average, the workstation processes the intermediate units more slowly than they are arriving. Unlike the case with $\rho > 1$, when $\rho < 1$ even as the buffer size tends to infinity, the marginal effect of the agent's effort on the probability of blocking does not go to zero, because no matter how large the buffer, the principal can never eliminate the threat of blocking. In this setting, the principal continues to increase the informativeness of throughput by increasing b , as this lowers the probability of blocking, which in turn allows the agent a greater opportunity to influence the probability of blocking, and consequently, throughput. This logic is summarized in the following corollary:

Corollary 1: When $\rho < 1$, the informativeness of throughput, K , with respect to the agent's effort, r , is everywhere increasing in the buffer size, b .

Because b is integer-valued, when the optimal buffer is finite (as when $\rho > 1$), it is difficult to characterize the optimal buffer in closed form. However, Proposition 3 establishes the interval in which the optimal b must lie as a function of ρ .

Proposition 3: When $1 < \rho < 2$, the optimal b lies in the interval: $\frac{1}{\sqrt{\rho-1}} - 1 < b \leq \frac{1}{\rho-1}$.

When $\rho \geq 2$, the optimal buffer is $b = 1$.

The first part of Proposition 3 indicates that the common objective in the JIT literature of eliminating WIP inventory buffers ($b = 1$) would incur agency costs if the

agent is not being asked to work relatively hard (i.e., $1 < \rho < 2$). More generally, our results point to the complementarities that are created between the choice of buffer size and the choice of production schedule, ρ , as a result of the effect of both on the informativeness of throughput as the performance metric.¹⁷ Changes in buffer size without associated changes in production schedule can lead to unintended agency costs.

Our analysis of the effect of buffer size on the informativeness of throughput is based on the buffer's role in blocking and starving the production process, with the other organizational design variables (r^*, λ, h) held fixed. However, we can also view these design variables in terms of their potential effects on the informativeness of throughput via their effects on the probabilities of blocking and starving. Situations can exist in which it may be more efficient to manipulate the informativeness of throughput by varying these other design choices than by varying the buffer size. For example, increasing buffer size has space utilization implications that the other variables do not; i.e., one cannot increase buffer size if there is no additional space available, whereas it may be possible to upgrade the workstation machinery (increase h) or speed up/slow down the upstream workstations (vary λ). Similarly, buffer size increments come in discrete steps while the other design variables may allow for finer adjustments. This motivates us to consider the effect of these other design choices on the informativeness of throughput while holding buffer size fixed, which is considered next.

Proposition 4: Holding the buffer fixed, the informativeness of K to r is:

¹⁷ Thus, our results are consistent with the complementarities literature, for example, (Milgrom and Roberts 1990) and (Milgrom and Roberts 1992).

- a. increasing in the mean arrival rate of intermediate units, λ ,
- b. decreasing in the principal's desired level of agent effort, r^* , and
- c. single-peaked in amount of help resources allocated to the agent, h .

Increasing λ decreases the likelihood of starving, which in turn reduces the effect of the variation in arrivals on throughput, making throughput more informative about the other source of variation, the agent's effort, r . The same logic holds for decreases in r . The more interesting result concerns productivity-enhancing investments, h . Proposition 4 implies that even ignoring out-of-pocket costs, there are two counter-acting effects to increasing productivity-enhancing investments for the agent. The direct positive effect is that such investments make the agent more productive, thereby increasing his ability to influence throughput, blocking and starving. The indirect, negative effect is that such investments may render the agent *so* productive, that holding effort fixed, the threat of blocking tends to zero as h increases, even if the agent's effort is only negligible. Because throughput only reflects effort to the extent that the agent's effort affects the probability of blocking beyond some point (1), increases in h make throughput less sensitive to the agent's effort and therefore less informative about the agent's effort, making it more costly to motivate him. In a sense, h plays the same role as buffer size in our earlier analysis. Both h and b can be looked at as productivity-enhancing investments and both can have negative agency cost effects as a result of their effects on the informativeness of throughput as a performance metric.

One can interpret Propositions 1, 3 and 4 as highlighting a value to creating bottlenecks, either by choosing a finite buffer size or a finite amount of productivity-

enhancing investments. The value arises for incentive reasons. This goes against conventional wisdom in the Operations Management literature, where the primary focus is on lost revenues rather than agency-cost minimization. Combining the two objectives into the firm's objective function may mitigate the force of our results; however because all of our findings are obtained using arbitrary parameter values, the incentive issues identified would persist in the face of additional objectives. Appendix A provides a numerical example for which we solve the principal's expected steady-state profit maximizing problem. The solution illustrates that the agency effect that we have found continues to hold.

III. Conclusion

As noted in the Introduction, the previous literature has largely ignored the effect of organizational design decisions on the informativeness of performance metrics. The present study analyzes the relation between the choice of the firm's production line buffers and optimal incentives. The Operation Management literature has focused on ways of reducing the probability of blocking and starving in production systems. A major insight from the present work is that we identify an incentive value for blocking and starving that arises from the firm's choice of finite buffers. We show that under reasonable conditions the informativeness of throughput as a performance metric is single-peaked in buffer size and therefore, it is optimal to induce non-zero probabilities of blocking and starving by setting the buffer size to an interior value ($\infty > b > 1$). We further show that the information-maximizing buffer is decreasing in the effort which the principal wants to induce from the agent. Therefore, we would expect to observe smaller buffer sizes in settings with more aggressive production schedules. Our findings thus emphasize the incentive complementarity between buffer size and production schedules, and provide a possible explanation for the mixed empirical evidence regarding the profitability of work-in-process inventory reductions (see (Nagar et al. 2009)).

This research spans two different disciplines, Managerial Accounting and Operations Management, each with its own nuances and paradigms. It represents an attempt to bridge these two literatures by weaving together some of the basic models of each. To do so, we have had to make a number of modeling assumptions. We ignored some of the factors that are identified in the Operations Management literature as influencing the inventory buffer decision in order to focus on incentive issues and the

informational effect of the buffer size decision. For example, some of the claimed benefits of small buffers under JIT which we have ignored are: the ability to identify production problems earlier, shorter lead-times, and faster responsiveness to changes in market conditions. However, we are confident that the incentive consequences of inventory buffers brought to light by our analysis will be present notwithstanding these other factors.

A specific limitation of our model is the assumption of linear compensation. While we made this assumption for analytical reasons, the crucial aspect for preserving our qualitative results is the monotonicity of compensation in the agent's marginal expected productivity, as is either assumed or shown to be optimal in the majority of agency models. Our analysis also assumes that the agent is risk-neutral. However, recall that the principal chooses b based on the information content of throughput \tilde{K} with respect to the agent's effort. Therefore, to the extent that a risk-averse agent must be compensated for risk, intuition would suggest that our results would be strengthened in the presence of risk-aversion. However, introducing risk-averse agents imposes several analytical complications, and we leave the resolution of such issues to future research.

Another simplifying assumption is that the agent's compensation is based on a single performance metric. Clearly, other performance metrics are possible, including realized inventory level.¹⁸ However, if mean steady-state inventory level is the performance metric, then it does not offer any information above and beyond the mean steady-state throughput, K , since both measures are functions of exactly the same variables. On the other hand, if we allowed for *continuous* monitoring, as with RFID,

¹⁸ We thank a referee for raising this issue.

then we could isolate the agent from the stochasticity of the arrival process and mitigate the agent's moral hazard problem.¹⁹ In this case, our problem reduces to the traditional moral-hazard setting where the agent always has work to process, and the principal infers the agent's effort via stochastic output. It would therefore be interesting to empirically analyze the incentive effect of buffers both before and after the installation of an RFID monitoring application to see the extent to which they are used as substitutes to filter noise in available performance measures. We leave to future research such inquiries, as compensation data surrounding RFID installations is not yet readily available.

¹⁹ Radio-frequency identification (RFID) involves "...an integrated circuit with an antenna, known as a 'tag,' attached to a...product. Product information as well as other relevant information can be stored in the tag. ... Using wireless technologies, readers can be set up to read the information on the tags..." (Lee and Ozer 2007, p. 40).

Appendix A

In this appendix, we provide a numerical example in which the principal's objective is to maximize her steady-state expected profits. Her choice variables are buffer size, b , the compensation contract, (w_0, w) , and the agent's effort level, r^* , where the latter is subject to moral hazard. Unlike the analysis in the text, this example does not take a Grossman-Hart approach but is instead a full expected profit maximization in which the principal chooses both the buffer size b and the effort level r^* to induce the agent to implement. We then compare our optimal buffer to that of the same problem except that the agent is not subject to moral hazard but the principal still chooses to induce the agent to implement the same effort level r^* . Comparing the two optimal solutions allows us to isolate the incentive effects of the principal's choice of buffer size. The example demonstrates that our findings continue to hold in the full, steady-state, profit-maximization problem. In particular, the incentive effect of buffers continues to exist and can bias the principal's profit-maximizing buffer away from the no moral hazard solution.

The example assumes that the principal's revenue per unit of throughput is R dollars. To facilitate the exposition, we fix $h = 0$. The principal's steady-state expected profit-maximization problem is thus:

$$\begin{aligned} & \text{Max}_{\{r, w, w_0, b\}} E_{\tilde{K}} R(\tilde{K}(b, r)) - wK(b, r) - w_0 - cb \\ & \text{s.t.} \\ & w\tilde{K} + w_0 - ar \geq 0 \quad \forall \tilde{K} \quad (LL) \\ & r \in \arg \max_r (wK(b, r) + w_0 - ar). \quad (IC) \end{aligned}$$

Varying the size of the buffer has three effects: it incurs a carrying cost of c per unit of capacity, it substitutes for effort in throughput, K , and it affects the agency

cost paid to the agent. The latter two effects directly influence the principal's preferred action choice. The relative importance of the three different effects varies with the specific parameters employed.

In our example, we set $R = 350$, $c = 10$, $a = 1$ and $\lambda = 1$. The principal finds it optimal to induce the agent to work relatively hard ($\rho = r^* = 2.11$) and to set the buffer size at $b = 3$. Figure A1 plots both the expected steady-state profits and the informativeness of throughput while fixing the induced effort at $r^* = 2.11$, but varying the buffer size. If there were no moral hazard problem and the principal set $r^* = 2.11$, the profit-maximizing buffer size would be $b = 4$, as indicated in Figure A1.

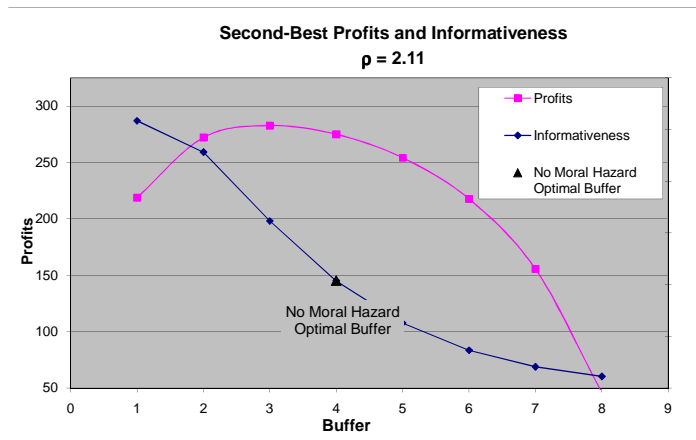


Figure A1: For $(R = 350, c = 10, a = 1, \lambda = 1)$, the moral hazard optimal solution is $\rho = r^* = 2.11$ and $b = 3$. The curves are drawn for fixed $\rho = r^* = 2.11$ but varying b . For the no moral hazard case in which the principal chooses $\rho = r^* = 2.11$, the optimal buffer size is $b = 4$.

The incentive effect of buffers is illustrated by the difference between the optimal buffer in the presence ($b = 3$) and absence ($b = 3$) of moral hazard on the part of the agent. Recall from Proposition 3 that when $\rho > 2$, the informativeness of throughput is always *decreasing* in buffer size, as in Figure A1. As a result, the principal finds it optimal to *reduce* her choice of buffer size from $b = 4$ to $b = 3$ in order to increase the informativeness of throughput and reduce the agency costs. However she doesn't reduce the buffer to its minimum, $b = 1$, because the resulting reduction in throughput and hence revenue, would more than offset the savings from reduced the agency costs.

Appendix B

Lemma 0: *The function $f \in C^k[a, b]$ has a root of multiplicity k at r if and only if:*

$$0 = f(r) = f'(r) = f''(r) = \dots = f^{(k-1)}(r) \quad \text{but} \quad f^{(k)}(r) \neq 0.$$

Proof of Lemma 0: See Theorem 3.20 and Corollary 3.21 in (Vinberg 2001) and

(Artin 1991). \square

Proof of Lemma 1: The agent's utility is given by $u = wK(b, r) + w_0 - ar$. Hence:

$$\frac{\partial u}{\partial r} = w \frac{\partial K}{\partial r} - a$$

$$\frac{\partial^2 u}{\partial r^2} = w \frac{\partial^2 K}{\partial r^2}.$$

Substituting the functional form of K from (1) yields:

$$\frac{\partial^2 u}{\partial r^2} = (w) \frac{\partial^2 K}{\partial r^2} =$$

$$(1+h)^2 (w) (b+1) \left(\frac{\lambda}{r(1+h)} \right)^{b+1} \frac{\left(b(r(1+h) - \lambda) \left(\lambda \left(\frac{\lambda}{r(1+h)} \right)^b + r(1+h) \right) + 2r(1+h) \lambda \left(\left(\frac{\lambda}{r(1+h)} \right)^b - 1 \right) \right)}{\left(\lambda \left(\frac{\lambda}{r(1+h)} \right)^b - r(1+h) \right)^3},$$

and the denominator is non-positive if and only if $r(1+h) \geq \lambda$. Further,

$$(w) (b+1) \left(\frac{\lambda}{r(1+h)} \right)^{b+1} > 0. \text{ Hence, if we can show that}$$

$\left(b(r(1+h) - \lambda) \left(\lambda \left(\frac{\lambda}{r(1+h)} \right)^b + r(1+h) \right) + 2r(1+h)\lambda \left(\left(\frac{\lambda}{r(1+h)} \right)^b - 1 \right) \right)$ is non-

negative if and only if $r(1+h) \geq \lambda$, we will then have established that $\frac{\partial^2 u}{\partial r^2} \leq 0$.

To this end, we can re-write the expression as:

$$(r(1+h))^{-b} \left[b(r(1+h))^{2+b} + (r(1+h))^{b+1} (-b\lambda - 2\lambda) + r(1+h)\lambda^{b+1}(2+b) - b\lambda^{b+2} \right],$$

labeling the expression in square brackets, N . According to Descartes' rule, N can have at most three roots in r . Using Lemma 0, note that:

$$N \Big|_{r=\frac{\lambda}{1+h}} = \frac{\partial N}{\partial r} \Big|_{r=\frac{\lambda}{1+h}} = \frac{\partial^2 N}{\partial r^2} \Big|_{r=\frac{\lambda}{1+h}} = 0 < b(1+h)^3 \lambda^{b-1} (2+3b+b^2) = \frac{\partial^3 N}{\partial r^3} \Big|_{r=\lambda},$$

which shows that all three roots are at $r(1+h) = \lambda$. Finally, since the third derivative

is positive, the Extremum Test assures that N has a saddle point at $r(1+h) = \lambda$;²⁰

therefore N alternates in sign as $r(1+h)$ crosses λ . However, since at $r(1+h) = 0$,

N is negative, N must be non-negative for $r(1+h) \geq \lambda > 0$.

The constraint (IC) can be expressed as:

$$\frac{\partial}{\partial r} \left(w \frac{1 - \left(\frac{r(1+h)}{\lambda} \right)^{-b}}{1 - \left(\frac{r(1+h)}{\lambda} \right)^{-(b+1)}} \lambda - ar \right) = 0,$$

implying,

$$w(b, r) = a \left(\frac{\partial K(b, r)}{\partial r} \right)^{-1} = \frac{a \left(\left(\frac{r(1+h)}{\lambda} \right)^{b+1} - 1 \right)^2}{(1+h) \left(1 - \left(\frac{r(1+h)}{\lambda} \right)^b \left(1 - b \left(1 - \frac{r(1+h)}{\lambda} \right) \right) \right)},$$

²⁰ For an explanation of the Extremum Test, see <http://mathworld.wolfram.com/ExtremumTest.html>

and w is both positive and increasing in r . \square

Proof of Proposition 1: Because we have defined informativeness as a constant times the inverse of the incentive wage, it suffices to show that the incentive wage is either “U” shaped in buffer size, or increasing in b . In particular, we will show that if the wage is increasing in b , at say \hat{b} , then the wage continues to increase for all $b > \hat{b}$. Then we will show that there exists a finite b beyond which the wage must be increasing in b .

In general, we can write:

$$w(b, \rho) = \frac{a(\rho^{b+1} - 1)^2}{(1+h)(b\rho^{b+1} - (b+1)\rho^b - 1)}, \quad (1)$$

where, henceforth to simplify the notation onwards, we omit the constant $\frac{a}{(1+h)}$.

Hence,

$$\frac{\partial}{\partial b} w(b, \rho) = \rho^b (\rho^{b+1} - 1) \frac{-(\rho - 1)(\rho^{b+1} - 1) + (b(\rho - 1)(1 + \rho^{b+1}) - \rho(\rho^b - 2) - 1)\log(\rho)}{(b\rho^{b+1} - (b+1)\rho^b + 1)^2}. \quad (2)$$

We first claim that the denominator of (2) is strictly positive. To see this, note that $b\rho^{b+1} - (b+1)\rho^b + 1$ has a root at $\rho = 1$; its first derivative evaluated at $\rho = 1$ is zero and its second derivative is strictly positive for all $\rho \geq 1$, implying that the denominator is strictly positive for all $\rho > 1$. Thus, the sign of $\frac{\partial}{\partial b} w(b, \rho)$ is determined by the function $t(b, \rho)$, where:

$$t(b, \rho) = -(\rho - 1)(\rho^{b+1} - 1) + (b(\rho - 1)(1 + \rho^{b+1}) - \rho(\rho^b - 2) - 1)\log(\rho).$$

We will show that if $t(b, \rho)$ is positive, it is increasing in b , implying that once the wage is increasing in buffer size b , it will continue to be increasing in b for larger buffers. To this end, note that the first term in $t(b, \rho)$ is strictly negative for $\rho > 1$. We first show that the second term in $t(b, \rho)$ is strictly positive for $\rho > 1$. To see this, note that we can write the coefficient of $\log(\rho)$ as

$-1 - b + (2 + b)\rho - (b + 1)\rho^{1+b} + b\rho^{2+b}$, which is zero at $\rho = 1$, strictly increasing in ρ at $\rho = 1$ and which has a strictly positive second derivative in ρ for $\rho > 1$:

$$b(1 + b)(\rho^{b-1}(2\rho + b(\rho - 1) - 1)) > 0.$$

Thus, if $t(b, \rho)$ is positive, it must be the case that:

$$\log(\rho) > \frac{(\rho - 1)(\rho^{b+1} - 1)}{b(\rho - 1)(1 + \rho^{b+1}) - \rho(\rho^b - 2) - 1}. \quad (3)$$

To show that if $t(b, \rho) > 0$, $t(b, \rho)$ is increasing in b , we difference $t(b, \rho)$ in b :

$$t(b + 1, \rho) - t(b, \rho) = (\rho - 1) \left(-(\rho - 1)\rho^{1+b} + (1 - (b + 1)\rho^{1+b} + (b + 1)\rho^{2+b}) \log(\rho) \right). \quad (4)$$

The coefficient of $\log(\rho)$ in (4) is always positive, hence we can use the bound for

$\log(\rho)$ from (3) to form a lower bound for $t(b + 1, \rho) - t(b, \rho)$:

$$\begin{aligned} t(b + 1, \rho) - t(b, \rho) &= (\rho - 1) \left(-(\rho - 1)\rho^{1+b} + (1 - (b + 1)\rho^{1+b} + (b + 1)\rho^{2+b}) \log(\rho) \right) \\ &> (\rho - 1) \left(-(\rho - 1)\rho^{1+b} + (1 - (b + 1)\rho^{1+b} + (b + 1)\rho^{2+b}) \frac{(\rho - 1)(\rho^{b+1} - 1)}{b(\rho - 1)(1 + \rho^{b+1}) - \rho(\rho^b - 2) - 1} \right) \\ &= (\rho - 1)^2 \frac{-1 + (3 + 2b)\rho^{1+b} - (3 + 2b)\rho^{2+b} + \rho^{3+2b}}{-1 - b + \rho(2 + b) - (b + 1)\rho^{1+b} + b\rho^{2+b}}. \end{aligned}$$

The denominator above is identical to that in (3), which we have already shown was strictly positive for $\rho > 1$. Thus we need only sign the numerator, which has at most three roots according to Descartes' rule. However, in accordance with Lemma 0, the numerator has all three roots at $\rho = 1$, and since it has a strictly positive third derivative in ρ , the expression is strictly for $\rho > 1$, implying that if at some $b = \hat{b}$ the wage is increasing in buffer size ($t(\hat{b}, \rho) > 0$), then it will continue to be increasing in buffer size for larger buffers.

To prove that the wage is eventually increasing in buffer-size, note that for $\rho > 1$, $\lim_{b \rightarrow \infty} t(b, \rho) = \infty$, implying that there exists a finite \tilde{b} for which $t(\tilde{b}, \rho) > 0$. \square

Proof of Proposition 2: In order to characterize the most informative buffer-size, it suffices to characterize the lowest incentive wage, w . We will show that if the wage is minimized at \hat{b} when $\rho = \bar{\rho}$, then the wage will be increasing in b at \hat{b} when $\rho > \bar{\rho}$; thus by Proposition 1, the wage minimizing buffer for $\rho > \bar{\rho}$ is less than or equal to \hat{b} .

Suppose the wage is minimized at \hat{b} when $\rho = \bar{\rho}$. Then by definition of a minimum: $w(\hat{b}, \bar{\rho}) - w(\hat{b} - 1, \bar{\rho}) \leq 0 \leq w(\hat{b} + 1, \bar{\rho}) - w(\hat{b}, \bar{\rho})$. To complete the proof, we will prove that $w(\hat{b} + 1, \bar{\rho}) - w(\hat{b}, \bar{\rho}) \geq 0$ implies $w(\hat{b} + 1, \rho) - w(\hat{b}, \rho) > 0$ for $\rho > \bar{\rho}$.

To this end, consider the difference $w(b + 1, \rho) - w(b, \rho)$:

$$(\rho-1)\rho^b \frac{1+b-3\rho(b+1)+3\rho^{2+b}+\rho^{3+b}-(2+b)\rho^{3+2b}+b\rho^{4+2b}}{(1-(b+1)\rho^b+b\rho^{1+b})(1-(2+b)\rho^{1+b}+(b+1)\rho^{2+b})}. \quad (5)$$

We first claim that the denominator of (5) is always positive. Using Descartes' rule, the first term has at most two roots. The term and its first derivative in ρ evaluated at $\rho = 1$ are both equal to zero. Therefore Lemma 0 implies that its only positive root is $\rho = 1$ and since its leading coefficient is positive, the term is strictly positive for $\rho > 1$. The same argument shows that the second term in the denominator is strictly positive for $\rho > 1$.

Thus, the sign of (5) is determined by the numerator. Applying Descartes' rule, the numerator has at most four positive roots in ρ . Because the numerator and both its first and second derivatives in ρ are equal to zero when $\rho = 1$, Lemma 0 implies that $\rho = 1$ is a root with multiplicity three. Because the third derivative of the numerator in ρ evaluated at $\rho = 1$ is negative $(-3b^2 - 9b - 6)$, but the leading coefficient is positive, the final root in ρ , which we label $\rho(b)$, must be greater than 1.

Now, suppose $w(\hat{b}+1, \hat{\rho}) - w(\hat{b}, \hat{\rho}) \geq 0$. Then it must be the case that $\hat{\rho}$ is greater than or equal to the final root, $\rho(\hat{b})$; therefore, if $\rho > \hat{\rho}$, then $\rho \geq \rho(\hat{b})$, in which case $w(\hat{b}+1, \rho) - w(\hat{b}, \rho) > 0$, and thus the wage-minimizing buffer for $\rho > \hat{\rho}$ must be no larger than \hat{b} . \square

Proof of Corollary 1: The proof is similar to that of Proposition 2. \square

Proof of Proposition 3: By definition, the information-maximizing b will equivalently minimize the wage, w . We identify a sufficient upper bound on b , as parameterized by ρ , for $w(b+1, \rho) - w(b, \rho)$ to be positive, and a sufficient lower bound on b for $w(b+1, \rho) - w(b, \rho)$ to be negative. Combining the two bounds, we identify a region for b where $w(b+1, \rho) - w(b, \rho)$ crosses zero in b , defining the wage-minimizing buffer size. We begin with two additional Lemmas:

Lemma 2: Let $f(x) = a_n x^n + \dots + a_1 x + a_0$ be a real polynomial with $f(1) = 0$.

Furthermore, assume that there exists a $k < n - 1$ such that for $i > k$, $a_i > 0$ and for $i \leq k$, $a_i \leq 0$. Then $f(x) > 0$ for $x > 1$ and $f(x) < 0$ for $x < 1$.

Proof of Lemma 2: For all $x > 1$ we have

$f(x) > (a_n + a_{n-1} + \dots + a_{k+1})x^{k+1} - (-a_k - a_{k-1} - \dots - a_0)x^k$. Since $f(1) = 0$, the sum of the terms in the parentheses are equal, and we denote the sum by s . Clearly, $s > 0$ and $sx^k(x-1) > 0$ for $x > 1$; hence, $f(x) > 0$ when $x > 1$. Similarly, for all $0 \leq x < 1$, we have $f(x) < sx^{k+1} - sx^k = sx^k(x-1) < 0$. This concludes the proof of the lemma. \square

Lemma 3: Let b be a natural number and $\rho > 1$, then:

$$\frac{2}{b}\rho + 1 \leq \frac{\sum_{i=0}^b (i+1)\rho^i}{\sum_{i=0}^{b-1} (i+1)\rho^i} \leq \frac{b+1}{b}\rho + \frac{1}{b}$$

for $\rho > 1$, and the opposite weak inequalities hold when $\rho < 1$.

Proof of Lemma 3:

First consider the inequality on the right-hand side. Set

$$f(p) = \left(\frac{b+1}{b}\rho + \frac{1}{b} \right) \sum_{i=0}^{b-1} (i+1)\rho^i - \sum_{i=0}^b (i+1)\rho^i.$$

We want show that $f(p) > 0$ for $p > 1$ and $f(p) < 0$ for $p < 1$. Thus, it is sufficient

to show that $f(p)$ satisfies the hypothesis of the previous lemma. First, to show that

$f(1) = 0$, we have:

$$\begin{aligned} f(1) &= \left(\frac{b+1}{b} + \frac{1}{b} \right) \sum_{i=0}^{b-1} (i+1) - \sum_{i=0}^b (i+1) \\ &= \left(\frac{b+2}{b} \right) \left(\frac{b}{2} \right) (b+1) - \frac{(b+1)(b+2)}{2} = 0. \end{aligned}$$

We must now find a k as in the prior lemma. If $f(p) = a_n p^n + \dots a_1 p + a_0$, then

$n = b - 1$ (although there is a p^b term in the construction of $f(p)$, note that it has a

coefficient of zero) and for arbitrary $m \geq 1$, the coefficient a_m is given by:

$$\frac{b+1}{b}m + \frac{1}{b}(m+1) - (m+1) = \frac{2m - (b-1)}{b}.$$

Hence, $a_m > 0$ if and only if $m > \frac{b-1}{2}$. In particular, a threshold k exists when $n = b - 1 > \frac{b-1}{2}$, which is valid for $b \geq 2$. If $b = 1$, then $f(p) = \frac{1}{2}(p - 1)$, which is greater than zero for $p > 1$ and less than zero for $p < 1$. Hence Lemma 2 holds and we have established Lemma 3 for the right-hand side inequality.

Next, consider the left-hand side inequality in the statement of the lemma. Let

$$g(p) = \sum_{i=0}^b (i+1)\rho^i - \left(\frac{2}{b}p + 1\right) \sum_{i=0}^{b-1} (i+1)\rho^i.$$

Note that:

$$g(1) = \sum_{i=0}^b (i+1) - \left(\frac{2}{b} + 1\right) \sum_{i=0}^{b-1} (i+1) = 0.$$

Moreover, the leading coefficient of $g(p)$ is positive for $b > 1$ and all the remaining coefficients are always negative; hence the conditions of Lemma 2 are always satisfied for $b > 1$. If $b = 1$, then $g(p) = 2(p - 1)$, which is again positive for $p > 1$ and negative for $p < 1$, as was to be shown. This concludes the proof of the lemma \square

Continuing with the proof of Proposition 3, we can rewrite $w(b, \rho)$ without

the constant $\frac{a}{(1+h)}$ as:

$$w(b, \rho) = \frac{(\rho^{b+1} - 1)^2}{(b\rho^{b+1} - (b+1)\rho^b + 1)} = \frac{\rho^{2b} + 2\rho^{2b-1} + \dots + (b+1)\rho^b + b\rho^{b-1} + \dots + 2\rho + 1}{b\rho^{b-1} + (b-1)\rho^{b-2} + (b-2)\rho^{b-3} + \dots + 2\rho + 1}.$$

To see this, we begin by rewriting the numerator:

$$\begin{aligned}
(\rho^{b+1} - 1)^2 &= ((\rho^{b+1} + \rho^b + \rho^{b-1} + \dots + \rho) - (\rho^b + \rho^{b-1} + \dots + 1))^2 \\
&= ((\rho - 1)(\rho^b + \rho^{b-1} + \dots + 1))^2 \\
&= (\rho - 1)^2 (\rho^b + \rho^{b-1} + \dots + 1)^2 \\
&= (\rho - 1)^2 \left(\sum_{i=0}^b \rho^i \right)^2 \\
&= \rho^{2b} + 2\rho^{2b-1} + \dots + (b+1)\rho^b + b\rho^{b-1} + \dots + 2\rho + 1.
\end{aligned}$$

Whereas the denominator can be rewritten as:

$$\begin{aligned}
b\rho^{b+1} - (b+1)\rho^b + 1 &= b(\rho^{b+1} - \rho^b) - (\rho^b - 1) \\
&= b\rho^b(\rho - 1) - (\rho - 1) \sum_{i=1}^b \rho^{b-i} \\
&= (\rho - 1) \left(b\rho^b - \sum_{i=1}^b \rho^{b-i} \right) \\
&= (\rho - 1) \sum_{i=1}^b (\rho^b - \rho^{b-i}) \\
&= (\rho - 1)^2 \sum_{i=0}^{b-1} (i+1)\rho^i \\
&= b\rho^{b-1} + (b-1)\rho^{b-2} + (b-2)\rho^{b-3} + \dots + 2\rho + 1.
\end{aligned}$$

In light of this representation, we can continue simplifying to obtain:

$$\begin{aligned}
w(b, \rho) &= \frac{\rho^{2b} + 2\rho^{2b-1} + \dots + (b+1)\rho^b + b\rho^{b-1} + \dots + 2\rho + 1}{b\rho^{b-1} + (b-1)\rho^{b-2} + (b-2)\rho^{b-3} + \dots + 2\rho + 1} \\
&= \frac{\rho^{2b} + 2\rho^{2b-1} + \dots + (b+1)\rho^b}{b\rho^{b-1} + (b-1)\rho^{b-2} + (b-2)\rho^{b-3} + \dots + 2\rho + 1} + 1 \\
&= \rho^{2b} \frac{1 + 2\rho^{-1} + \dots + (b+1)\rho^{-b}}{b\rho^{b-1} + (b-1)\rho^{b-2} + (b-2)\rho^{b-3} + \dots + 2\rho + 1} + 1. \quad (6)
\end{aligned}$$

Using (6), $w(b+1, \rho) > w(b, \rho)$ is equivalent to the following equivalent inequalities:

$$\begin{aligned}
& \rho^{2b+2} \frac{1+2\rho^{-1}+\dots+(b+2)\rho^{-b-1}}{(b+1)\rho^b+\dots+2\rho+1} + 1 > \rho^{2b} \frac{1+2\rho^{-1}+\dots+(b+1)\rho^{-b}}{b\rho^{b-1}+\dots+2\rho+1} + 1 \\
& \Leftrightarrow \rho^2 \frac{1+2\rho^{-1}+\dots+(b+2)\rho^{-b-1}}{(b+1)\rho^b+\dots+2\rho+1} > \frac{1+2\rho^{-1}+\dots+(b+1)\rho^{-b}}{b\rho^{b-1}+\dots+2\rho+1} \\
& \Leftrightarrow \rho^2 \left(\frac{(b+1)\rho^b+\dots+2\rho+1}{1+2\rho^{-1}+\dots+(b+1)\rho^{-b}} \right) \left(\frac{1+2\rho^{-1}+\dots+(b+2)\rho^{-b-1}}{(b+1)\rho^b+\dots+2\rho+1} \right) \\
& \quad > \left(\frac{(b+1)\rho^b+\dots+2\rho+1}{1+2\rho^{-1}+\dots+(b+1)\rho^{-b}} \right) \left(\frac{1+2\rho^{-1}+\dots+(b+1)\rho^{-b}}{b\rho^{b-1}+\dots+2\rho+1} \right) \\
& \Leftrightarrow \rho^2 \frac{1+2\rho^{-1}+\dots+(b+2)\rho^{-b-1}}{1+2\rho^{-1}+\dots+(b+1)\rho^{-b}} = \rho^2 \frac{\sum_{i=0}^{b+1} (i+1)\rho^{-i}}{\sum_{i=0}^b (i+1)\rho^{-i}} > \frac{(b+1)\rho^b+\dots+2\rho+1}{b\rho^{b-1}+\dots+2\rho+1} = \frac{\sum_{i=0}^b (i+1)\rho^i}{\sum_{i=0}^{b-1} (i+1)\rho^i}. \quad (7)
\end{aligned}$$

We can use Lemma 3 with $\rho > 1$ on the right-hand side of (7) to conclude that it is less than or equal to $\frac{b+1}{b} \rho + \frac{1}{b}$. Since we have assumed that $\rho > 1$, $\rho^{-1} < 1$ and we can again apply Lemma 3, replacing ρ with ρ^{-1} on the left-hand side of (7), to conclude that it is greater than or equal to $\rho^2 \left(\frac{b+2}{b+1} \rho^{-1} + \frac{1}{b+1} \right)$. Thus a sufficient condition on b and ρ for $w(b+1, \rho) > w(b, \rho)$ is given by:

$$\begin{aligned}
& \frac{1}{b+1} \rho^2 + \frac{b+2}{b+1} \rho > \frac{b+1}{b} \rho + \frac{1}{b} \\
& \Leftrightarrow \frac{1}{b+1} \rho^2 + \frac{b+2}{b+1} \rho - \frac{b+1}{b} \rho - \frac{1}{b} > 0.
\end{aligned}$$

Solving the second quadratic above for equality gives two roots in ρ : -1 and $\frac{b+1}{b}$.

Since the leading coefficient is positive, we know that (7) holds whenever $\rho > \frac{1+b}{b}$, or

a sufficient condition for the wage to be increasing in b is given by $b > \frac{1}{\rho-1}$.

Next we establish a sufficient upper bound for the wage to be decreasing in b .

When $a > 0$ and $\rho > 1$, $\rho^{-a} < 1$; thus Lemma 3 implies that the left-hand side of (7)

is less than $\rho^2 \left(\frac{2}{b+1} \rho^{-1} + 1 \right)$, whereas the right-hand side is bounded below by $\frac{2}{b} \rho + 1$.

Hence a sufficient condition for (7) to be violated (for the wage to be decreasing in b)

is given by:

$$\frac{2}{b+1} \rho + \rho^2 - \frac{2}{b} \rho - 1 < 0. \quad (8)$$

The two roots to the polynomial on the left-hand side of (8) are given by

$$\rho = \frac{\frac{2}{b(b+1)} - \sqrt{\frac{4}{b^2(b+1)^2} + 4}}{2} \quad \text{and} \quad \rho = \frac{\frac{2}{b(b+1)} + \sqrt{\frac{4}{b^2(b+1)^2} + 4}}{2},$$

since the latter is negative and the polynomial

has a positive leading coefficient; a sufficient condition for the wage to be decreasing in b is given by $1 < \rho < \frac{\frac{2}{b(b+1)} + \sqrt{\frac{4}{b^2(b+1)^2} + 4}}{2}$. A lower bound for $\frac{\frac{2}{b(b+1)} + \sqrt{\frac{4}{b^2(b+1)^2} + 4}}{2}$ is given by

$$\frac{\frac{2}{b(b+1)} + 2}{2} = 1 + \frac{1}{b(b+1)},$$

and since $b(b+1) < (b+1)^2$, a sufficient condition for $w(b, \rho)$ to be decreasing is thus $\rho < 1 + \frac{1}{(b+1)^2}$, or $\frac{1}{\sqrt{\rho-1}} - 1 > b$. \square

Proof of Proposition 4:

In order to prove that the sensitivity of K to r is increasing in λ , decreasing in r , and single-peaked in h , we will show that the wage w is decreasing in λ , increasing in r and “U” shaped in h . Differentiating w with respect to h yields:

$$\frac{\partial w}{\partial h} = a \frac{\left(-1 + \rho^b (1 + 2b + b^2) - \rho^{b+1} (4b + b^2) + \rho^{2b+2} (1 + 4b) - \rho^{3b+2} (1+b)^2 + \rho^{3b+3} b^2 \right)}{\left(\rho^2 (1 + \rho^b (\rho b - b - 1)) \right)^2}. \quad (9)$$

The denominator is unambiguously non-negative, whereas Descartes’ rule implies the numerator has at most five positive roots. Because the numerator and the first three derivatives with respect to ρ have a root at $\rho = 1$, only one root remains unidentified.

If $b = 1$, then the fourth derivative of the numerator in (9) is zero at $\rho = 1$, and thus

the fifth root is in fact $\rho = 1$. If $b > 1$, then the fourth derivative of the numerator in

(9) is strictly positive at $\rho = 1$, hence by the Extremum Test, $\rho = 1$ is a relative minimum and there exists a sufficiently small $\varepsilon > 0$ such that the numerator is positive at $\rho = 1 - \varepsilon$ when $b > 1$. However, because the numerator in (9) is strictly negative at $\rho = 0$, the final root must belong to the interval (0,1).

Differentiating w with respect to r yields:

$$\frac{\partial w}{\partial r} = -a \frac{(1+b)\rho^{b-1}(1-\rho^{b+1})(-b+\rho(b+2)-\rho^{b+1}(b+2)+\rho^{b+2}b)}{\lambda(1-\rho^b(1+b)+b\rho^{b+1})^2}. \quad (10)$$

The denominator of (10) is non-negative. Descartes' rule implies that in addition to $\rho = 1$, (10) may have up to three additional positive roots. Using Lemma 0, the expression $-b + \rho(b+2) - \rho^{b+1}(b+2) + \rho^{b+2}b$ and its first two derivatives have a root at $\rho = 1$, whereas the third derivative is strictly positive at $\rho = 1$, implying that the expression has a saddle point at $\rho = 1$. Since the pre-multiplier, $1 - \rho^{b+1}$, also alternates in sign at $\rho = 1$, the entire expression, $\frac{\partial w}{\partial r}$, is non-positive. Finally,

differentiating w with respect to λ yields: $\frac{\partial w}{\partial \lambda} = -\frac{r}{\lambda} \frac{\partial w}{\partial r}$. \square

References

- Alles, M., S. Datar, and R. A. Lambert. 1995. Moral Hazard and Management Control in Just-in-Time Settings. *Journal of Accounting Research* 33 (Supplement):177-204.
- Artin, M. 1991. *Algebra*. Upper Saddle River, N.J.: Prentice-Hall.

- Arya, A., and J. Glover. 2008. Performance Measurement Manipulation: Cherry-Picking What to Correct. *Review of Accounting Studies* 13 (1):119 - 139.
- Balachandran, K., and S. Radhakrishnan. 1996. Cost of Congestion, Operational Efficiency and Management Accounting. *European Journal of Operational Research* 89 (2):237 - 245.
- Balakrishnan, R., N. J. Nagarajan, and K. Sivaramakrishnan. 1998. The Effect of Property Rights and Audit Information Quality on Team Incentives for Inventory Reduction. *Management Science* 44 (9):1193 - 1204.
- Balsamo, S., V. Personé, and R. Onvural. 2001. *Analysis of Queueing Networks with Blocking*. Boston: Kluwer Academic Publishers.
- Banker, R., and S. Datar. 1989. Sensitivity, Precision and Linear Aggregation of Signals for Performance Evaluation. *Journal of Accounting Research* 27 (1): 21-39.
- Berg, N., and N. Fast. 1975. The Lincoln Electric Company. Harvard Business School Case 9-376-028: Harvard Business School.
- Berkley, B. 1992. A Review of the Kanban Production Control Research Literature. *Production and Operations Management* 1 (4):393 - 411.
- Crémer, J. 1995. Towards an Economic Theory of Incentives in Just-In-Time Manufacturing. *European Economic Review* 39 (3-4):432-439.
- Datar, S., and M. Rajan. 1995. Optimal Incentive Schemes in Bottleneck-Constrained Production Environments. *Journal of Accounting Research* 33 (1):33-58.
- Demougin, D., and D. Garvie. 1991. Contractual Design with Correlated Information under Limited Liability. *Rand Journal of Economics* 22 (4):477 - 489.
- Dutta, S., and F. Gigler. 2002. The Effect of Earnings Forecasts on Earnings Management. *Journal of Accounting Research* 40 (3):631 - 655.
- Feltham, G., and J. Xie. 1994. Performance Measure Congruity and Diversity in Multi-Task Principal/Agent Relations. *The Accounting Review* 69 (3):429-453.
- Gietzmann, M., and T. Hemmer. 2002. On the Relation Between Optimal Incentive Structures and the Cost and Benefits of Bottlenecks. *Journal of Labor Economics* 20 (2, part 2):34 - 56.
- Gigler, F., and T. Hemmer. 2002. Informational Costs and Benefits of Creating Separately Identifiable Operating Segments. *Journal of Accounting & Economics* 33 (1):69 - 90.
- Groenevelt, H. 1993. The Just-In-Time System. In *Handbooks in OR & MS*, edited by S. Graves, A. Rinnooy Kan and P. Zipkin. Amsterdam: Elsevier, 629 - 670.
- Grossman, S., and O. Hart. 1983. An Analysis of the Principal Agent Problem. *Econometrica* 51 (1):7-45.
- Hall, B., C. Madigan, and E. Lazear. 2000. Performance Pay at Safelite Auto Glass. Harvard Business School Case 800291: Harvard Business School.
- Hemmer, T. 1995. On the Interrelation Between Production Technology, Job Design, and Incentives. *Journal of Accounting and Economics* 19 (2-3):209-245.
- . 1998. Performance Measurement Systems, Incentives, and the Optimal Allocation of Responsibilities. *Journal of Accounting & Economics* 25 (3):321-347.

- Holmstrom, B. 1979. Moral Hazard and Observability. *Bell Journal of Economics* 10 (1):74-91.
- Holmstrom, B., and P. Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design. *Journal of Law, Economics, and Organization* 7 (0):24-52.
- Hopp, W., and M. Spearman. 2001. *Factory Physics*. 1 ed. Chicago, IL: Irwin.
- Kim, S.-H., M. Cohen, and S. Netessine. 2007. Performance Contracting in After-Sales Service Supply Chains. *Management Science* 53 (12):1843 - 1858.
- Lee, H., and O. Ozer. 2007. Unlocking the Value of RFID. *Production and Operations Management* 16 (1):40 - 64.
- Lu, L., J. Van Mieghem, and R. Savaskan. 2006. Incentives for Quality through Endogenous Routing. Center for Operations and Supply Chain Management, Kellogg School of Management, Northwestern University.
- Melumad, N., D. Mookherjee, and S. Reichelstein. 1995. Hierarchical Decentralization of Incentive Contracts. *The Rand Journal of Economics* 26 (4):654 - 672.
- Milgrom, P., and J. Roberts. 1990. The Economics of Modern Manufacturing: Technology, Strategy, and Organization. *American Economic Review* 80 (3):511-528.
- . 1992. *Economics, Organization & Management*. Englewood Cliffs, NJ: Prentice Hall.
- . 1995. Complementarities and Fit: Strategy, Structure, and Organizational Change in Manufacturing. *Journal of Accounting & Economics* 19 (2-3):179 - 208.
- Nagar, V., M. Rajan, and R. Saouma. 2009. The Incentive Value of Inventory and Cross-training in Modern Manufacturing. *Journal of Accounting Research* 47 (4):991-1025.
- Radhakrishnan, S., and K. Balachandran. 1995. Delay Cost and Incentive Schemes for Multiple Users. *Management Science* 41 (4):646 - 652.
- . 2004. Service Capacity and Incentive Compatible Cost Allocation for Reporting Usage Forecasts. *European Journal of Operational Research* 157 (1):180 - 195.
- Riordan, M., and D. Sappington. 1987. Information, Incentives, and Organizational Mode. *Quarterly Journal of Economics* 102 (2):243-263.
- Vinberg, E. B. 2001. *A Course in Algebra*. Vol. 56. Providence, R.I.: American Mathematical Society.