Abstract Title: Examination of the Use of Variability Buffers in Healthcare

Sriram Venkataraman
Department of Management, 401 Sirrine Hall
Clemson University, Clemson, SC 29634. Email: svenkat@clemson.edu

Lawrence D. Fredendall
Department of Management, 123B Sirrine Hall
Clemson University, Clemson, SC 29634. Email: flawren@clemson.edu

Kevin M. Taaffe
Department of Industrial Engineering, 130A Freeman Hall
Clemson University, Clemson, SC 29634. Email: taaffe@clemson.edu

Nathan Huynh
Department of Civil and Environmental Engineering, 300 Main Street, Room C116
University of South Carolina, Columbia, SC 29208. Email: huynhn@cec.sc.edu

POMS 22nd Annual Conference
Reno, Nevada, U.S.A.
April 29 to May 2, 2011
1 Introduction

This paper, which is a case study of a Perioperative Services (POS) department of a major teaching hospital in the southeast, evaluates the operation of a POS in terms of existing operations models. This paper makes three contributions to our understanding of healthcare services. First, it generalizes the model of a manufacturing system as four variables - three buffers and system variance (Hopp and Spearman 2004) – to healthcare services. Second, it extends the methodology developed by Olivares et al. (2008) to examine how operating room managers make scheduling decisions to examine decisions about the entire POS. Third, it enriches the model proposed by Hopp and Spearman (2004) by integrating variables proposed by Hayes et al. (1988) to achieve a richer understanding of the decisions managers make about the POS in particular and healthcare services in general.

Generalizing the model proposed by Hopp and Spearman (2004) for manufacturing to healthcare services is an important contribution to our ability to understand this complex service process. Hopp and Spearman (2004) model firms as having three buffers of inventory, capacity and lead time that protect the system from variability. By using this four variable model we are able to investigate how decisions are made within this very complex healthcare service system. While all of the interactions of these four variables within the POS are not fully understood, the focus in this research on just these four variables provides researchers with a path forward which will provide insight and point towards potential improvements in decision making. This research seeks to determine if this model from manufacturing is directly applicable to services. A major difference between manufacturing and services is that services such as POS have many elements which are not tangible and are not easily quantified (e.g., Fitzsimmons and Fitzsimmons 2008). In addition, Hopp et al. (2007) identified situations in services where there are dis-
cretionary task completion times that influenced both service quality and service capacity.

Enriching the model proposed by Hopp and Spearman (2004) using the variables proposed by Hayes et al. (1988) discussion of how firms achieve world class manufacturing status was drawn on to investigate the paths by which managers create the three buffers in the POS service system and how their decisions may reduce or amplify the variance in the system. This paper contributes to theory development by using this case study to understand why managers make certain decisions that affect these variability buffers and to examine the consequences of these decisions. By doing so, this case study places the concepts of Hopp and Spearman (2004) within the context of service operations strategy.

This paper not only advances research into healthcare systems, it also furthers the development of theory in operations management. While there has been much research about operations there has been limited theory development. For example, Schmenner and Swink (1998) proposed the theory of swift-and-even flow but did not provide causal direction between the variables. In their four variable model, Hopp and Spearman (2004) demonstrate that these variables form a reciprocal model only. By integrating this model with the framework provided by Hayes et al. (1988) we are able to suggest causal linkages between the variables over time.

This paper extends the methodology developed by Olivares et al. (2008). In their research, Olivares et al. (2008) analyzed reservations of operating room times using structural estimation technique from econometrics and they empirically verified the validity of the famous “newsvendor problem” to this decision. We confirm that this technique measures managers’ perception of the relative costs of overtime to undertime in a POS. Further, it supports other data about the size of the three buffers in the POS.
The rest of this paper is organized in the following way: Section 2 is literature review, section 3 explains data, methodology, and hypothesis, section 4 covers results and discussion, and section 5 includes conclusions, and limitations.

2 Literature Review

Three streams of research are reviewed here. First, the research that places the work of Hopp and Spearman (2004) within the broad context of work design within the operations management field is discussed. Second, the use of the newsvendor model within analytical models in operations research is discussed and then the application of this model by Olivares et al. (2008) is explained in detail. Finally, the stream of literature related to Hayes et al. (1988) framework about managerial decision making is discussed.

Work design is a larger construct than job design. Work design includes all of the systems and processes used to organize the work in an organization including the support services, while job design is concerned about the individual jobs (Sinha and de Ven, 2005). Work design includes the allocation of tasks and responsibilities and the design of systems to ensure effective communication and integration of the various tasks and responsibilities (Daft and Lengel, 1986). Sinha and de Ven (2005) argue that work design is of fundamental importance because it "directly affects the behavior and performance of individual workers and organizations each day, as well as the aggregate productivity and well-being of economies " (p. 389). They classify work design issues into three interrelated problem: 1) modularity problem of dividing work between organizational units; 2) a hierarchical problem of coordinating work and responsibilities across hierarchical levels; and 3) a network problem of the interactions of the vertical
and horizontal divisions of work and responsibilities.

Much of the work design research conducted in operations has been concerned with the issues of coordinating work both vertically and horizontally. One significant stream is that of workload control (see Fredendall et al. 2010 for a complete review) which limits the amount of work that is within the system at any point in time. Hopp and Spearman (2004) argue that this limit on the work in the system ensures that the system is easier to control than those systems without a limit on the work. Hopp and Spearman (2004) expand on this by conceptualizing any operations system as a set of three interrelated buffers to protect the system from variability. The three buffers are an inventory buffer, a capacity buffer and a lead time buffer. The interrelations of these variables is seen through Little’s Law. For example, if the inventory buffer is increased by allowing more work into the system then according to Little’s Law the lead time is simultaneously increased unless the system throughput is increased by increasing the capacity buffer.

Because decisions about each buffer are made at different time points, Hopp and Spearman (2004) say that many managers do not understand the interactions of these decisions on these buffer levels. For example, if a manager wants short lead times for customers, the manager will increase overtime if the lead times grow too large. This overtime increases the capacity buffer which reduces the lead time (i.e., reduces the lead time buffer). As the lead-time gets smaller, the inventory buffer becomes smaller and the manager may respond by reducing the size of the capacity buffer.

The amount of variance in a system and the buffers in a system are the result of managerial decisions. Hayes and Wheelwright (1984) and Hayes et al. (1988) argued that effective execution of strategy
requires that managers understand how their decisions about infrastructure variables influences their operations capability. [Hayes et al. (1988)] listed these infrastructure variables as human resource policies and practices, quality assurance and control system, production planning and inventory control system, new product development processes, performance measurement and reward systems and organizational structure and design. The relationships between their decisions and how these decision categories as a framework to discuss the decisions that firms make to create the variability buffers and how these decisions interact with each other. The key decisions about the capacity and inventory buffers are analyzed using techniques developed by [Olivares et al.] (2008).

Multiple authors have demonstrated the relationship between variability and the work-in-process inventory (e.g., [Anupindi et al. 2005]) and the time for a job to serviced (e.g., [Hopp et al. 2007]) by using variations of the Pollaczek-Khintchine formula ([Heyman and Sobel] 1982). Indeed, [Lovejoy (1998)] suggests that these queuing relationships could form the foundation of a theory of operations management that will allow findings from other theories to be integrated. A moderating influence on the use of queuing theory to explain why operations perform as they do is the task environment As the need for problem solving increases there is the need for more decentralized decision-making, which requires different ways of coordinating labor resources. In a complex system managers must coordinate multiple resources to achieve their desired outcomes. Work design typically divides tasks to create specialized labor resources, but then requires mechanisms to reintegrate or coordinate these specialized resources ([Lovejoy 1998]). A key responsibility of managers is to coordinate the multiple resources used to accomplish work. The level of coordination required depends on the level of variance and the demands on the resources.

There have been many papers in the Operations Research and Operations Management literature
that have analytically used the “newsvendor model”. Newsvendor model is used when the demand is a random variable. Using the notations by Olivares et al. (2008), let D be a random variable with a known distribution \( F(\cdot) \) and let Q be the decision made by the newsvendor before D’s realization takes place. If \( Q > D \), then the newsvendor incurs a cost of \( C_o(Q - D)^+ \) and if \( D > Q \), then the newsvendor incurs a cost of \( C_u(D - Q)^+ \); where \( C_o \) and \( C_u \) are overage and underage cost parameters respectively, and \( C_o \) and \( C_u \) > 0; and \( (z)^+ = \max\{z,0\} \). The objective of the newsvendor problem is to find the optimal \( Q^* \) that minimizes the expected total cost, i.e.

\[
Q^* = \arg\min_Q E[C_o(Q - D)^+ + C_u(D - Q)^+]
\]  

It has been established (see Zipkin, 2000) “that the optimal solution to (1) satisfies” (Olivares et al., 2008):

\[
F(Q^*) = \frac{C_u}{C_u + C_o} = \frac{1}{1 + \gamma}
\]  

where \( \gamma = C_o/C_u \) is the cost ratio. \( \gamma \) is used instead of the explicit costs \( C_o \) and \( C_u \) to avoid identification problems later in the analysis.

Most of the existing papers have used the newsvendor problem in analytical settings, where (generally) the distribution function \( F(\cdot) \) of the random variable D is known, cost parameters \( C_o \) and \( C_u \) are known, and the optimal value of Q (\( Q^* \)) is computed. However, in this paper we use the econometric framework developed by Olivares et al. (2008) in which we assume the decision-maker (newsvendor) is rational (so, \( Q^* \) is known), and the the distribution function \( F(\cdot) \) of the random variable D is known, and we estimate the cost ratio (\( \gamma \)) empirically using data. Olivares et al. (2008) formulated a econometric framework and tested it empirically in a hospital setting. In particular, the authors estimated the cost ratio involving the costs of over-reserving versus under-reserving an operating room. The authors found many schedule-overruns and concluded that the hospital gave more importance to not paying over-time,
although they were fine with paying idle-time costs. The authors also included forecasting-biases (from the medical literature) in their model. [Cohen et al., 2003] studied the order-fulfillment process of a company that is producing customized devices. Since these are customized devices like defense equipments, medical devices, etc., the company producing such equipments (hereon referred to as the supplier) usually receives a “soft order” consisting of the shared forecasts. This helps the supplier in deciding when to start fulfilling the orders given a “stochastic internal manufacturing lead time” (p. 1653, Cohen et al., 2003). This decision of when to start fulfilling the orders is similar to a newsvendor problem since if the supplier starts fulfilling the orders too early, then it will incur an extra holding cost; whereas if the supplier starts fulfilling the orders too late, then it will incur a delay cost. Based on the assumption that the supplier acts “rationally” the authors found that the supplier views the “holding cost to be about three times higher than the delay cost” (p. 1653, Cohen et al., 2003).

3 Data, Methodology, and Hypothesis

The first step in the analysis was to study the POS in detail to understand the system. Staff in each of the three POS departments were shadowed by members of the research team and time studies of each activity performed at each step in the POS’s set of activities were conducted. The researchers followed the staff member for a period of 2 to 8 hours depending on the repetition in their activities. These were used to create process flow maps of each unit in the POS to identify work flows, the sources of variability and the coordination mechanisms used to achieve coordination at each stage of the POS. The technique of shadowing to gather detailed data about the actual activities of a job has been used by many researchers (e.g., Fredendall et al., 2009, Yule et al., 2006). The flow of patients within the POS is from Preoperative (Pre-Op), to the Operating Room (Intraoperative) to Post Anesthesia Care Unit (PACU). The supporting
departments schedule the OR, supply medicine, supply sterile supplies and sterilize instruments after surgeries. It was determined that instrument sterilization, pharmacy and sterile supplies were not affecting OR operations and they were not studied further.

Figure 1: Perioperative Services: Patient and Supporting Flows

The second step in the analysis was to evaluate the OR performance using archival data. Ten days of data from June 2010 were used to calculate the average utilization, the total on-time starts and the First-Case on-time starts. During this week there were 660 cases scheduled to begin and end between 8 AM and 6 PM. This time period is the busiest time period during the 24 hour period. While each of these 660 cases was a different patient, there may have been multiple procedures performed on the same patient during the scheduled time period. However, any surgeons performed multiple surgeries on different patients during a day.
This POS had 30 OR at that time, so there were \(10 \times 30 \times 10 = 3,000\) hours available to schedule during the two weeks. The schedule for the 660 cases used 1300.3 hours of this available including scheduled room turn around activities. This is a 43.34\% utilization. Of these 660 cases only 9 started on-time for a 1.36\% average. Of the 150 cases scheduled to start at 8 AM, only 5 had started by 8:30 AM for a 3.33\% First-Case-On-Time start rate.

The next step in the analysis was to examine three months’ worth of scheduling for the operating room to evaluate the managerial decisions. We had three months of data for operating room (from June 2010 to August 2010). We used the econometric framework developed by Olivares et al. (2008) in this paper. So, we assumed that \(D_i\)s are drawn independently from \(F(\cdot, \theta)\), where \(\theta \in \Theta\) is a parameter vector and \(\Theta\) is a parameter space. Following the notations used by Olivares et al. (2008) and the functional econometric form, let

\[
\theta_i = h(X_i, \eta) \tag{3}
\]

where \(X_i\) represents the covariates and \(\eta\) represents the estimator vector. Similarly, let:

\[
\gamma_i = g(Z_i, \alpha) \tag{4}
\]

where \(Z_i\) represents the covariates and \(\alpha\) represents the estimator vector. We could estimate \(\alpha\) using linear regression, however \(\gamma_i\) is not unknown.

Using the framework developed by Olivares et al. (2008) and equations 2, 3, and 4, we get:

\[
F(Q^*_i; h(X_i, \eta)) = \frac{1}{1 + g(Z_i, \alpha)} \tag{5}
\]

Consistent with the medical literature, we assumed log-normal distribution for cost ratio. So,

\[
log(\gamma_i) = Z_i \alpha + \psi_i . \tag{6}
\]
Since $\gamma_i$ is not observed, we need to use an estimate of $\gamma_i$ (see [Murphy and Topel, 2002]). So, we use the following 2-step procedure as proposed by [Olivares et al., 2008] and [Murphy and Topel, 2002]:

1. First, we estimate $\hat{\theta} = h(X_i, \hat{\eta})$

2. Second, we estimate $\hat{\gamma} = \frac{1}{F(Q_i; \hat{\theta}_i)} - 1$ and finally estimate $\alpha$ by $\hat{\alpha}$ using equation 6.

For operating room, we used surgeon experience level, charge level (acuity level), case type code (type of surgery), and multproc (whether surgery involved multiple procedures or not) as covariates for both stages of estimation procedure. Since maximum likelihood technique and ordinary least squares (OLS) will produce same estimates (see [Ruud, 2000]), we used OLS to get the estimates of $\eta$. We had 4588 observations initially. We deleted 123 observations since these corresponded to acuity level: “fixed fee” and the supervisor of pre-op told us that “fixed-fee” could correspond to any level of acuity. Further, we deleted 5 observations since we did not have data on surgeons’ experience level for those observations.

### 3.1 Hypothesis

We developed a hypothesis based on the framework developed by [Hopp and Spearman, 2004] and tested it using the methodology developed by [Olivares et al., 2008]. We expect $\hat{\gamma}$ to be less than one for operating room the over-time costs of operating room staff are much higher than idle-time costs. So, formally our first hypothesis is stated as follows:

**Hypothesis 1:** For operating room, $\hat{\gamma} < 1$.

This hypothesis is well supported by the past medical literature (see [Strum et al., 1999]) as well as the results obtained by [Olivares et al., 2008].
4 Results and Discussion

The first step in the analysis demonstrated that the workload into the system is controlled. While there are some emergency surgeries that are scheduled as needed, the OR scheduling department levels the number of surgeries that occur on a daily basis. Given Hopp and Spearman’s (2004) argument that the workload must be controlled to have stable buffers it is expected that the buffers in the OR should be stable. There was more capacity in the preoperative department and the PACU than in the operating rooms. The Pre-Op department has enough capacity to have the patients prepared well in advance of their scheduled surgery.

In the week of observing the Pre-Op department there were no instances where the start of a surgery was delayed because of a scarcity of capacity in Pre-Op. All of the delays were due to late arrival of lab reports or the surgeons or anesthesiologists were delayed in visiting the patient to mark the site of the surgery or to update the Health and Physical report.

Similar observations of the PACU during this time period showed that it had more capacity than the OR. This protective capacity meant that the PACU was able to accept patients from the OR with less than 15 minutes notice. The staff reported that the PACU had not stopped an OR in the last five years of POS operation. To do this the POS uses extra capacity from Pre-Op as flexible, surge capacity for PACU. The Pre-Op nurses have the ability to move from Pre-Op to PACU and the supervisory staff can also work with patients if required.

The POS does use two types of inventory buffers. First it uses patients who are prepared to go into the OR as a buffer. While many patients wait in the Pre-Op past the scheduled start of surgery, others will start earlier if there is room in both the surgeon’s and OR schedules. The second POS inventory
buffer is instruments and sterile supplies. These are typically prepared and in the OR core by 3 AM of the morning of the surgery. The surgical technician can then quickly setup the instruments in the room. However, there may be a shortage of some special equipment such as portable x-ray machines that could delay the surgery.

The POS uses two lead-time buffers. Patients are frequently requested to move their scheduled surgery forward if the POS scheduling office recognizes that they have some opening in the OR schedule. To do this the POS scheduling office contacts surgeons and patients to move surgeries forward. Once inside the POS the patients who are in the inventory buffer described above are told that their surgery will start in 150 to 400 minutes.

The POS also uses a capacity buffer. As calculated earlier, it has a 43.34% utilization. To reduce undertime costs, if there are no problems, the POS managers send staff home on flex time before the end of their scheduled shift. If overtime is needed to complete an OR schedule then it is used.

To examine the thoughts of the managers about overtime versus undertime, the OR data was analyzed using Olivares et al. (2008) method. To model the first stage of estimation of operating room, we regressed:

\[
\ln(D) = X\beta + \epsilon
\] (7)

where D is a vector of actual surgery time and X is the matrix of covariates [surgeon experience level, charge level (acuity level), case type code (type of surgery), and multproc (whether surgery involved...
multiple procedures or not). Then, we used the estimates of $\hat{\beta}$ to calculate

$$\hat{\gamma}_i = \frac{1}{\Phi\left(\frac{\ln(Q_i) - X_i\hat{\beta}}{\hat{\sigma}_i}\right)} - 1$$

Then, we regressed:

$$\ln(\hat{\gamma}) = X\alpha + \psi.$$  \hspace{1cm} (9)

We deleted 61 more observations since they had standardized residuals greater than 3 or less than -3.

Table 1 shows the results of regression from step 1 estimation procedure. $R^2 = 0.3944$, with $p < 0.001$; so the covariates capture around 40% of variance in the actual time of surgery. This result is consistent with that reported by [Olivares et al. (2008)], although they only used cardio-vascular surgeries. Further, all the four covariates were found to significantly affect actual surgery time at 99%. As the experience level of surgeon increases, actual surgery time goes down (which makes sense and agrees with medical literature). Conversely, as acuity level goes up, actual time of surgery goes up (which again makes sense since higher acuity levels are associated with complex surgeries). Actual time of surgery is also higher for surgeries that involve multiple procedures than those that do not (consistent with the findings reported by [Olivares et al. (2008)]).

Table 2 shows the summary statistics of $\hat{\gamma}$. The median value of $\hat{\gamma}$ is 0.8794. This result is consistent with the median value of $\hat{\gamma}$ obtained by [Olivares et al. (2008)]. Further, this result confirms the first hypothesis and hence the hospital indeed pays more attention to over-time costs as opposed to idle-time costs in case of operating room.
5 Conclusions and Limitations

This paper extended the work by Olivares et al. (2008) to the entire peri-operative services unit of a leading hospital in South Carolina. Further, we tied up the results found by using a 2-step estimation procedure with buffering theory (which has been used a lot in Operations Management). This is one of the first attempts (atleast to our knowledge) to tie structural estimation procedure of econometrics with buffering theory of Operations Management.

We showed that actual surgery time can be predicted using four covariates (with a pretty good $R^2$). Further, all the four covariates were found to significantly affect actual surgery time. Charge level (acuity level) is one of the most important predictors and it was not used in the previous study by Olivares et al. (2008) since they just studied cardiovascular surgeries, whereas we had an independent variable “type of surgery”.

Since hypothesis 1 is supported by our data, this indicates that managers do avoid overtime and are willing to accept the costs of undertime. This agrees with the existence of a capacity buffer in the OR. Since each OR is associated with high labor costs (e.g., at a minimum 1 CRNA, 1 Circulating Nurse and 1 Surgical Technician at approximately $150 per hour) and an MD Anesthesiologist is assigned to multiple OR, undertime is a significant cost, but is considered to be much smaller than overtime.

An unresolved issue at this point is that utilization is only 43.34% but First-Case-On-Time-Starts is only 3.33% during the 10 day period examined. This does not seem to be due to surges in cases, since the scheduling office pulls cases forward to keep a level daily load. Shadowing the staff through Pre-op
suggested that most patients are ready to be transported to the OR at least 30 minutes before the scheduled start of the surgery except when there is a need for a health and physical update or the surgical site marking by the surgeon or anesthesiologist.

This analysis suggests that the low utilization of the OR is not due to external variance since the OR scheduling office attempts to level the workload. The analysis also suggests that the low utilization is not due to the Pre-Op or PACU departments since they are not constraints on patient preparation, nor do they restrict the flow of patients out of the OR.

Further research is needed to determine the sources of the variance that require such expensive capacity buffers. Further examination is also needed to see why the cheaper Pre-Op and PACU buffers are not more effective in keeping the OR fully utilized.
References


### Appendix

Table 1: Summary results of Step-1 estimation procedure of operating room

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 3.2821   | 0.0370     | 88.65   | 0.0000   |
| SurgExp        | -0.0063  | 0.0009     | -7.26   | 0.0000   |
| CSCode         | 0.0055   | 0.0011     | 4.84    | 0.0000   |
| Acuity         | 0.3751   | 0.0085     | 43.99   | 0.0000   |
| Mult.Proc      | 0.3330   | 0.0158     | 21.07   | 0.0000   |

Table 2: Summary Statistics of $\hat{\gamma}$ of operating room

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4399</td>
<td>0.8794</td>
<td>0.00122</td>
<td>342.22</td>
<td>3.5397</td>
<td>13.05956</td>
</tr>
</tbody>
</table>