Lean beyond waste

Roemeling, Oskar-Paul

CHAPTER 5

This chapter is currently under review at *Production Planning & Control* as: Roemeling, O.P., Land, M.J., & Ahaus, C.T.B. (2016). Buffering by adjusting processing times in healthcare.

Buffering by adjusting processing times in healthcare

5.1 INTRODUCTION

This research investigates the interactions between buffers and variability in healthcare environments from a Lean perspective. We are especially interested in the concept of quality buffers (Hopp *et al.* 2007). Quality buffers are identified as an addition to the widely recognized buffering opportunities in service environments. We focus on the theoretical implications of quality buffers and research the feasibility of applying this buffer type as well as the more commonly identified time and capacity buffers.

Through activities such as benchmarking or strategies aimed at improving patient flow (Cohen *et al.*, 2008), healthcare processes are continuously evaluated and improved (Umble and Umble, 2006). Improvements in patient flow have particularly been linked to increased healthcare quality and productivity (Litvak, 2009; Villa *et al.*, 2009). Limited patient flow in healthcare providers can arise from issues such as admitting unfit patients, lengthy discharge procedures, or poor coordination with external parties (Rich and Piercy, 2013). Flow can be improved by reducing

activities that do not add value, bottlenecks, and variability (Schmenner and Swink, 1998; Schmenner, 2001). The Lean philosophy is a popular strategy to improve process flow. Lean is a management system that aims to achieve the desired production with the minimum use of buffering (Hopp and Spearman, 2004) and is facilitated through a relentless process of continuous improvement (Shah and Ward, 2007; Hopp, 2008; LaGanga, 2011). Lean was initially seen as an appropriate approach for manufacturing, with studies showing promising results such as improved overall performance (Shah and Ward, 2003) and increased quality (Oliver *et al.*, 1994; Cua *et al.*, 2001). In this study, we are interested in Lean in the context of healthcare, and particularly in the roles of variability and buffers.

Whilst Lean is certainly popular in services (e.g. Piercy and Rich, 2009; Radnor and Johnston, 2013), including in healthcare (e.g. Mazzocato *et al*, 2010), it originated in a production environment where variability led to buffers in the forms of time, capacity, and inventory (Hopp and Spearman, 2008; Thürer *et al.*, 2014). In service environments, the service is produced and consumed simultaneously (Radnor and Osborne, 2013). Consequently, inventory buffers play a minor role in healthcare: the patients themselves are transformed in the process, and these patients cannot be stocked in advance of demand. This implies that coping with variability in healthcare and similar service environments will be regulated using time and capacity buffers (Jack and Powers, 2004). Capacity buffers present themselves in the form of idle capacity. In healthcare, one could think of physicians waiting for a patient so that they can provide treatment

or perform a diagnosis. Similarly, time buffers appear in the form of waiting patients: waiting for treatment, for diagnosis, or for other actions. In this research, we define capacity buffering in healthcare as the mechanism in which capacity resources are idle in order to absorb variability in the availability of patients or in the resource itself. We define time buffering as the mechanism in which patients are waiting in order to be able to absorb the variability in the availability of resources or in the patients themselves. The interaction between capacity and time buffers results in a continuous 'buffering trade-off' between additional waiting time for a patient and idle capacity. It seems that, when dealing with variability, healthcare providers have a choice between Scylla and Charybdis.

The ability to trade time and capacity buffers fails to fully explain the many situations in which patients are not waiting but capacity resources (such as physicians) are fully occupied. It seems that another factor is in play, and a possible explanation is provided by Hopp *et al.* (2007) who introduced an additional buffering mechanism, specific to service environments, that has been typified as a quality buffer. Hopp *et al.* (2007) argue that, in environments where actors have discretion over the completion of their task, these actors can increase or decrease the quality of the delivered service in response to potential idle capacity or imminent congestion. Other studies have acknowledged the possibility of quality buffers (e.g. Wang *et al.*, 2010; Kostami and Rajagopalan, 2013; Kuntz *et al.*, 2014), but we are unaware of any studies that have further explored quality buffers and investigated how they absorb variability. The central research questions for our

study were therefore: (1) how should the mechanisms behind quality buffers be typified and explained, and (2) how do quality buffers interact with time and capacity buffer mechanisms? The paper now continues with a critical reflection on the theory and the concept of Lean in relation to healthcare and the roles of variability and buffers within this. Following this, by combining theory and logic, we explain the relationship between buffers and the mechanisms behind quality buffers. We draw the paper to a close by drawing conclusions from our findings.

5.2. THEORY: THE INTERACTION OF VARIABILITY AND BUFFERS

The potential benefits of Lean approaches in healthcare look promising, with several studies reporting positive outcomes (Nelson-Peterson and Leppa, 2007; Raab *et al.*, 2008; Hydes *et al.*, 2012). Nevertheless, literature reviews struggle to provide a conclusive verdict on the effects of Lean in healthcare, with Brackett *et al.* (2013) unable to conclude that Lean greatly influences patient care. Further, Costa and Filho (2016) report that Lean in healthcare environments is still applied in a superficial way. In addition, Mazzocato *et al.* (2010) found that Lean-related studies in healthcare often report on the reduction of direct waste, suggesting that reducing variability and buffers receives less attention. This is perhaps surprising given that it is especially the reduction of variability that improves flow (Schmenner and Swink, 1998; Fredendall *et al.*, 2009; Drupsteen *et al.*, 2013). Variability is a major concern for hospital performance (Salzarulo *et al.*, 2011) since it puts a strain on capacity, and inadequate staffing levels

have been shown to result in increased patient mortality risk and increased staff turnovers (Aiken *et al.*, 2002).

In healthcare, one can distinguish between natural and artificial variability (Litvak and Long, 2000). Natural variability is a result of differences between individual patients and between medical staff, and cannot, or only to a limited extent, be influenced (Litvak and Long, 2000; Joosten *et al.*, 2009). Contrarily, artificial variability follows from controllable, non-random, factors and these can potentially be influenced and the variability reduced. In other words, our own actions are the underlying causes of artificial variability. Both types of variability lead to the use of time or capacity buffering, or a combination of the two. This implies idle capacity resources and/or patients waiting to receive attention. In both cases, one party is waiting for the other to become available. However, the introduction of quality buffers, identified by Hopp *et al.* (2007), may provide a third buffering opportunity.

Hopp *et al.* (2007) describe how quality buffers come about as follows: in systems with discretionary task completion, such as healthcare, it is attractive to complete a service earlier in response to congestion. Instead of using waiting as the buffer mechanism, quality can serve as a buffering mechanism. Quality buffers assume that service providers are able to adjust delivered service quality in order to cope with variability (i.e. offer a lower service quality when busy). In this perspective, the quality of a service is seen as the consequence of processing time adjustment rather than being directly affected. However, we would argue that a degraded service quality does not automatically follow from shortened processing times. Therefore, we would propose seeing

adapting the processing time as the buffering mechanism and therefore, instead of the term quality buffers, we refer to processing time (PT) buffers. In contrast to both time and capacity buffers, waiting is not a fundamental component of PT buffers. Since waiting is not fundamental to PT buffers, the question becomes how such buffers cope with variability. Here, we discern two possible perspectives in that PT buffers can:

(1) reduce or avoid future customer waiting time (i.e. time buffers) by decreasing processing time;

(2) reduce or avoid future idle resources (capacity buffers) by increasing processing time.

These two alternatives show parallels with inventory buffers in production settings. In a production setting, inventory buffers anticipate future demand which would otherwise result in backlogs (time buffers), and avoid idle capacity resources (capacity buffers) by producing ahead of demand. Similar to inventory buffers, PT buffers can work when it is not possible or desirable to keep a customer (i.e. patient) waiting or have idle excess capacity.

To provide a better understanding of PT buffers, we have investigated the interchange or 'trade-off' among the various buffer types. It is customary to view buffers as communicating vessels. Figure 5.1 depicts the trade-off between buffers in a service environment. In this situation, inventory is not considered to be a primary buffer type, and PT buffers may take the role over. Nevertheless, the basic premise remains: that if one buffer is increased, another buffer (or combination of buffers) should decrease. If we now consider a practical example where an emergency care unit is confronted with a sudden rise in the

number of patients, and all the medical and nursing staff are already fully occupied, but, somehow, the new patients do not experience longer waiting times before receiving treatment. To understand this situation, we need to incorporate PT buffers in our thinking as only then can we explain the stable waiting times. With all other aspects unchanged, a reduction in processing times is the only feasible explanation.

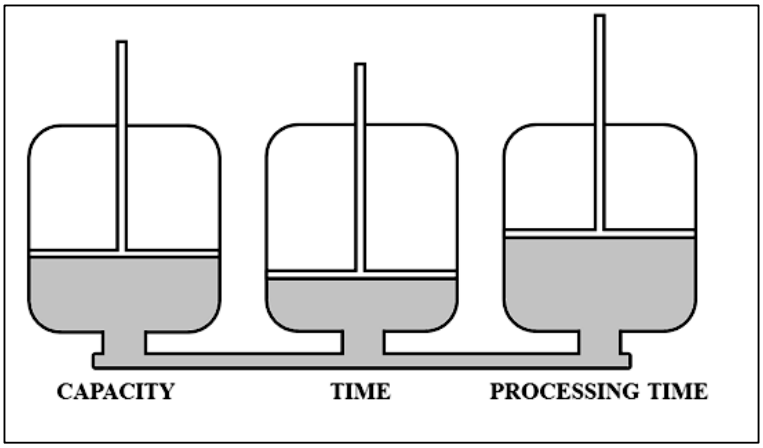

Figure 5.1. Representation of the relationship between buffers.

We have argued that quality buffers are more appropriately described as PT buffers since the prime adaptation mechanism relates to processing time, although this may indeed influence quality. However, we have not yet addressed how processing times can be reduced. In an empirical study, Batt and Terwiesch (2012) focused on service times in an emergency care unit and found that care providers are able to adjust their clinical behavior to increase service speed. One of the ways that medical professionals can increase service speed is to reduce service content. Essentially,

Batt and Terwiesch (2012) identified an opportunity to reduce processing time by changing the clinical behavior to reduce the number of tasks. In another study, Kc and Terwiesch (2009) investigated the effect of system load on service speed and found that workers who encountered a busy period, in terms of an increase in system load, increased their task speed. So, instead of added capacity or additional patient waiting times, the speed of the task is increased. The increased workload, which results in reduced processing times, is not without it risks and can lead to adverse quality effects (Oliva and Sterman, 2001). Based on the above examples, we can identify two options in the healthcare literature in order to adjust processing times:

(1) reduce or increase the number of tasks;

(2) reduce or increase task speed.

Clearly, cutting processing times by reducing the number of tasks or by increasing task speed can have consequences for service quality. However, we would argue that this is not always the case. It is difficult for patients to evaluate the technical skills of medical professionals and also the results of treatment (Kang and James, 2004). Patients therefore often rely on other measures of quality associated with the delivery of care. From a patient's perspective, attributes such as empathy can determine care quality (Kang and James, 2004). Thus, as long as attributes such as empathy are present, reducing processing time does not necessarily result in lower perceived quality. As such, the consequences of PT buffers, through sometimes reducing processing time, for quality are strongly dependent on how quality is defined. For example, if quality is seen as conformance to

specifications (Reeves and Bednar, 1994), then a physician's small talk ('bedside manner') might not contribute to quality viewed as the quick and full recovery of the patient. However, if quality is understood as exceeding customer expectations, a short conversation with a patient is probably part of the desired experience and therefore boosts quality. This also underlines why we see the original term 'quality buffer' as an unfortunate choice in this situation. The impact of adapting processing time is fully dependent on the way quality is perceived.

Once an interpretation of quality has been defined, one can distinguish between tasks that facilitate service quality, and those that do not. In other words, we can distinguish between critical and non-critical tasks. Critical tasks are those activities that cannot be accelerated or left out without affecting service quality. Non-critical tasks are activities that can be omitted without reducing service quality. As long as professionals focus on carrying out the critical tasks while reducing processing time, there will be no tangible, or at least measurable, implications for the health-related outcomes within the applied quality perspective.

Finally, we should consider whether such processing time adjustments should really be seen as a form of flexible capacity buffer since it is the capacity resource that determines which tasks are left out or sped up. However, we would argue that processing time adjustments should not be confused with flexible capacity. Hopp and Spearman (2008) argue that flexibility is an attribute of buffers. Buffer flexibility is defined as the possibility to use the same buffer for different purposes, leading to a reduction in overall buffering requirements (Iravani *et al.*, 2005; Hopp and

Spearman, 2008; Hopp and Lovejoy, 2013). In this respect, PT buffers can reduce the amount of buffering sourced from other buffer types, but they do not reduce the overall amount of buffering. Further, it is not a case of overcapacity being used for different purposes, as it is with flexible capacity, since PT buffers use the same amount of capacity but at different intensities. As such, we do not see capacity resources as the buffering mechanism although they do play an important role in applying PT buffers.

5.3 CONCLUSIONS

In this research, we set out to explore the merits of quality buffers in a healthcare environment and our main research questions were: (1) how should the mechanisms behind quality buffers be typified and explained, and (2) how do quality buffers interact with the more commonly identified time and capacity buffers? In response to our first research question, we showed that quality buffers could be better characterized as processing time (PT) buffers. The mechanism behind adjustments in processing times amounted to reducing the number of tasks and/or increasing task speed. These actions might result in reduced service quality but this is heavily dependent on the way quality is defined in a specific environment. In response to our second research question, we showed that PT buffers interact with other buffers in a service environment and work in a somewhat similar way to inventory buffers in a production situation. When one type of buffer is reduced, another buffer has to increase unless the underlying variability is reduced. Here, PT buffers help explain situations

where both capacity and patient waiting times remain stable when the variability increases.

Based on the preceding considerations, we would suggest the following definition of PT buffering: *"Processing time buffering is the mechanism through which processing times are reduced by removing critical and/or non-critical tasks and/or increasing task speed in response to variability"*. Quality is incorporated in this definition through the distinction made between critical and non-critical tasks. However, we argue that quality changes are a possible consequence of the processing time buffering mechanism, not the buffering mechanism itself. The consequences for quality of applying processing time buffers are far from straightforward, and depend on how quality is perceived.

This study is not without its limitations. First, in this study, we have relied on logic and examples taken from the literature. To boost confidence in the findings, controlled experiments could be used to establish if quality buffers are indeed enacted through adjusting both task speed and the number of tasks. There may also be other mechanisms in play. Nevertheless, this study does provide a stronger theoretical foundation for the existence and role of PT buffers, and this research explains why the alternative of thinking in terms of quality buffers is inappropriate in some situations. Future research could further explore the interchange between PT buffers and the time and capacity buffers identified in the literature. For example, it would be valuable to empirically establish the relationships between the buffer types in a range of service environments.