Working with capacity limitations: operations management in critical care

Christian Terwiesch, PhD[1], Jeremy M. Kahn MD MSc[2], Diwas KC, PhD[3]

1. Department of Operations Management, The Wharton School; Leonard Davis Institute of Health Economics, University of Pennsylvania
2. Clinical Research, Investigation and Systems Modeling of Acute Illness (CRISMA) Center, Department of Critical Care Medicine, University of Pittsburgh School of Medicine; Department of Health Policy and Management, University of Pittsburgh Graduate School of Public Health
3. Department of Information Systems and Operations Management, Goizueta Business School, Emory University

Author addresses: Christian Terwiesch, PhD (corresponding author)
Huntsman Hall 573
3730 Walnut Street
Philadelphia, PA 19104
Phone: 215.898.8541
Email: terwiesch@wharton.upenn.edu

Jeremy M. Kahn, MD MSc
Scaife Hall Room 602-B
3500 Terrace Street
Pittsburgh, PA 15261
Phone: 412.647.3136
Email: kahnjm@upmc.edu

Diwas KC, PhD
1300 Clifton Road NE
Atlanta, GA 30030
Phone: 404.727.1424
Email: diwas_kc@bus.emory.edu

Word count: 2825

SCENARIO

As your hospital's intensive care unit (ICU) director you are approached by the hospital's administration to help solve ongoing problems with ICU throughput. The ICU seems to be full all the time, and trauma patients in the emergency department sometimes wait up to 24 hours before receiving a bed. At the same time, the cardiac surgeons were forced to canceled several elective coronary-artery bypass graft cases because there was not a bed available for post-operative recovery. The hospital administrators ask if you can decrease your ICU length of stay, and wonder if they should build more ICU beds instead. For help in understanding and optimizing your ICU's throughput, you seek out the operations management researchers at your university.

INTRODUCTION

Increasing demand for critical care has made capacity limitations commonplace in the ICU [1]. These limitations occur when there are no available ICU beds for patients with critical illness, leading to delays in ICU admission that have important clinical and economic consequences. Admission delays can result in the boarding of critically ill patients in the emergency department or in other hospital units, which has been associated with increased risk of morbidity and mortality [2, 3]. Admission delays can also result in decreased revenue for hospitals, which depend on available beds to perform elective surgery cases or accept patients in transfer from outside hospitals. These problems have forced the critical care community to develop innovative ways to improve throughput and address capacity constraints. Yet these problems are not unique to the ICU, or even unique to health care in general. Limited capacity and the resulting problems of waiting times and throughput losses exist in many processes, ranging from financial services to automotive production. The academic field of operations management is specifically designed to address these issues. The purpose of this review is to provide a brief overview of operations

management and to present a set of case studies from work environments other than hospitals, thereby exposing readers to operations management and its potential application to critical care.

WHAT IS OPERATIONS MANAGEMENT?

Working with Capacity Limitations

Many operations, in particular service processes such as restaurants and airlines, have high fixed costs. These fixed costs typically reflect the cost of maintaining a certain capacity available, where capacity is defined as the maximum number of customers that can be served per unit of time. Often, these fixed costs are the wages required to pay labor or the cost of machinery for production. Yet while costs in services tend to be fixed, revenue increases proportionally to the number of customers served per unit time, also referred to as throughput. This scenario creates an economic incentive to operate the process at a high level of utilization, where utilization is defined as the ratio of the number of customers served (the throughput) to the maximum number of customers that we could serve (the capacity).

Consider the following simplified example. A service has a fixed cost of $1000 per day and obtains $20 per customer served. The operation thus breaks even at 50 customers served per day. At 60 customers per day, the service obtains $200 in profits per day. At 70 customers, the process obtains $400 in profits. In other words, increasing the number of customers served from 60 to 70 (a 10/60=16.7% increase) leads to a 100% increase in profits. The marginal (additional) cost of service is zero while the marginal revenue is high. Maximizing utilization becomes a key priority.

Understanding the problem

By definition, utilization cannot exceed 100%. Thus, the money seeking manager is tempted to seek near 100% utilization. High utilization, in and by itself, is not a problem. To see this, assume that in

an example process, customers would arrive exactly one customer every 5 minutes (12 customers arrive per hour). Further, assume that it takes us exactly 4 minutes to serve each customer (thus, we could serve up to 15 customers per hour). The resulting utilization in this process would be 12/15=80%. We might be tempted to call this a 20% under utilization and seek additional demand to improve our profitability.

However, this ignores an important reality of service delivery—variability. Customers are not widgets in an assembly line and the amount of time that it takes to serve one customer depends on the customer at hand. Furthermore, the arrival times of individual customers may not be known in advance. These sources of uncertainty create a stochastic effect on our process. Consider the data shown in Figure 1. Just as before, 12 customers arrive per hour. This time, however, the exact arrival times are random. Similarly, we again take 4 minutes, on average, to serve a customer. Yet, some customers get served quickly while others take longer. Although the mean demand and capacity remain constant, Figure 1 reveals that what previously appeared as an underutilized process is in reality a rather busy place. Indeed, some customers (e.g. the 5th and the 6th customer) spend much more time waiting than they spend in service. We also observe that the number of customers in the process at any one time goes up all the way to four (three waiting, one being served). Contrast this with the previous "deterministic" scenario, where each customer is served immediately upon arrival. Customers in the deterministic system incurred zero wait prior to initiation of service.

Variability is the enemy of operations. An 80% utilization of an automated assembly line with limited or no variability might be under-utilized; an 80% utilization of a time critical service in presence of variability is asking for trouble. The example in Figure 1 assumed that customers, in

cases of a temporary capacity shortfall, would patiently wait in line until it is their turn to be served. But it is easy to conceive of settings in which customers might not be able or willing to wait.
The branch of operations management that mathematically analyzes the interplay between process flows, utilization, and variability is referred to as Queuing Theory. Various mathematical models exist to inform the capacity planning in such an environment. For example, one might ask for the amount of capacity that is needed (the number of people to be hired, or equipment to be purchased) so that customers get served in a given expected wait time.

One of the most prominent findings in this line of work is the insight that the average waiting time increases dramatically at higher levels of utilization. Specifically, the average waiting grows proportionally to utilization/(1-utilization). This finding has substantial practical implications. For example, for a utilization of 80%, the ratio of 0.8/(1-0.8) equates to 4. For a utilization of 90%, this ratio grows to 0.9/(1-0.9)=9. Thus, a 10% increase in utilization can more than double the waiting time. This detrimental effect on the process's responsiveness needs to be kept in mind when we accept more demand in an attempt to increase utilization. Similar mathematical models exist for the case in which waiting is not possible. For example, one can predict the percentage of customers that will be lost due to capacity shortfalls, either because of customer impatience, or simply due to a lack of sufficient waiting space.

Better, not more

The above example illustrates a fundamental trade-off between the efficiency of a process as measured by its utilization and its responsiveness as measured by its wait time (Figure 2, top). Waiting time is reduced as more resources are added. Operations management tools, in particular queuing theory, can help to find the right positioning along the efficiency-responsiveness frontier. But Operations Management can do more than just trade-off one desirable process characteristic

against another. Operations Management is also about innovation. By creating an innovative process redesign, the aim is to shift out the frontier instead of simply supporting the optimal position on the current frontier (Figure 2, bottom). The process becomes better.

New frontiers might be reached by overcoming inefficiencies in the present process design (often referred to as waste) or by creating flexibility in the process to better cope with variability. For example, industrial pioneers such as Henry Ford redefined the production of physical goods. As work was increasingly divided, craftsmen were replaced by less skilled workers. Production processes were perfected over the subsequent decades, culminating in the legendary Toyota Production System (TPS) that is now widely regarded as the gold standard for excellent operations [4, 5]. TPS emphasizes the need to continuously improve a process, driving out the so-called seven sources of waste: excess production, waiting times, transport steps, excessively long activity times, inventory, rework (fixing quality problems), and unnecessary motions. Work flows are optimized, capacity levels are chosen so to match demand, activities are standardized, and protocols are implemented to standardize work, reduce defects, and improve productivity.

Example 1: Focus—the United States Airline industry and the emergence of Southwest

The US Airline industry is a tough place to compete in and many airlines have gone through extended periods of financial losses and bankruptcies. An interesting exception is Southwest Airlines, which has created a number of efficiency related innovations in the air travel process and in turn has been rewarded with outstanding growth and profitability. Many of these innovations reflect the company's decision to focus on specific market segments and operational processes. For example, Southwest offers only economy class seating, has a standardized check-in process, flies only one type of aircraft, and minimizes extraneous amenities such as meals and entertainment. Such focus has led to substantial process improvements, by reducing both customer-related

variability and process-related variability. Consequently, Southwest can achieve high levels of utilization *and* improved service times. Southwest has been able to obtain higher employee productivity (more passengers served per employee) and higher capital productivity (fewer airplanes per passenger due to fast gate turnaround times) while being able to command only marginally lower fares compared to their competitors (Figure 3).

Example 2: Quick response—local production and quick replenishment at Zara

Few industries are plagued by variability like the fashion industry. Consumer tastes are fickle and hard to predict. Most apparel retailers order their merchandise from East-Asia and typically must commit to orders a full year before the beginning of their season. Consequently, they often end up with too little of those products that later on emerge as the "hot" items of the season (leading to missed sales opportunities) and too much of other products (requiring substantial mark-downs). Mark-downs of 30-70% are common, especially towards the end of the season, and many stores spend as much as half of their store area on such unprofitable items. Zara's operational innovation has been one of local production, with approximately 50% of its merchandise sourced from its home country of Spain. At first glance, local production appears inefficient as wages in Spain are significantly higher than in East-Asia. However, the local production allows a quick and frequent replenishment and enables a tight integration between Zara's retail operation and their production process. As a result, Zara builds in flexibility into its operation and is able to react to unanticipated swings in demand as opposed to relying on an order for the entire selling season based on a single forecast.

Example 3: Capacity pooling and chaining—Honda's platform strategy

Another response to variability is based on aggregating its multiple sources. From basic statistics, we know that the risks associated with variability decrease as we aggregate many independent

sources of variability. For example, the financial risk of fire for an individual home owner is large, yet an insurance company with a million fire policies faces relatively lower risk. Aggregating variability across independent sources is the idea behind capacity pooling. Consider an automotive company that operates multiple plants manufacturing and produces different models. A given car model can only be produced in exactly one plant. If demand increases relative to the forecast, that plant is unlikely to have sufficient capacity to fulfill it. Conversely, if demand decreases, the plant is likely to have excess capacity. The company can mitigate some of the demand-supply mismatch by pooling its capacity. Specifically, if every model could be made at every plant, high demand from one model can be served with spare capacity due to low demand from another one, leading to better plant utilization and more sales. However, such capacity pooling would require the plants to be perfectly flexible – requiring substantial investments in production tools and worker skills. An interesting alternative to such perfect flexibility is the concept of partial flexibility, also referred to as chaining. The idea of chaining is that every car can be made in two plants and that the vehicle-to-plant assignment creates a chain that connects as many vehicles and plants as possible. It can be shown that such partial flexibility results in almost the same benefits of full flexibility, yet at dramatically lower costs [6]. Figure 4 provides an illustrative comparison of partial flexibility and full flexibility.

APPLYING OPERATIONS MANAGEMENT TO CRITICAL CARE

The capacity problems facing ICUs are directly analogous to our examples. The vast majority of critical care costs are fixed, resulting in substantial revenue increases with each additional patient [7]. ICUs also frequently operate at or near capacity, with subsequent increases in wait times [8]. Simply expanding capacity is not feasible, due to space limitations within hospitals, workforce shortages, and government regulations [9]. Neither is expanding capacity necessarily desirable. As

the above examples teach us, in the face of variable demand, expanding capacity will result in higher fixed costs, excess capacity and long-term inefficiencies.

The science of operations management is specifically designed to address these problems. ICU throughput is at heart a complex service problem—patients are just customers arriving at random times and with varying needs. Each takes a different amount of service time. The overall goal is to maximize quality while minimizing waste. In the ICU, quality comes in form of low mortality and waste comes in the form of wait times (i.e. admission delays), excess activity times (i.e. long lengths of stay), and the need for rework (i.e. the effort required to fix with ICU-acquired complications and readmissions). Operations management can help us not only trade-off capacity and efficiency under our current process but also help us "shift the frontier" through continuous process improvement.

The first step is to understand the current process. What is the ICU utilization how much does it vary? What are sources of ICU demand, and how much of that demand is random versus predictable? What is the average ICU length of stay (i.e. service time) and how does it differ between different patient types? How much of the current activity is true production versus waste in the form of ICU readmissions or discharge delays? The next step is to apply queuing theory to mathematically formulate the current process and determine the point on the utilization curve which will maximize responsiveness and productivity. Increasing capacity might be necessary to achieve optimal throughput, or might only result in excess resources.

The final step is the search for ways to improve the current processes to increase throughput with the current resources. Taking a lesson from Toyota, standardizing care through protocols might lead to decreased waste in the form of hospital-acquired infections or excess ventilator days [10]. Splitting the single surgical ICU into two subspecialty ICUs (one for trauma and one for cardiac

surgery) might introduce economies of scope, by which the specialty ICUs can perform their services more efficiently. This situation would be analogous to Southwest Airlines, which increased efficiency in part by limiting the scope of their services. To prevent adverse effects from boarding and to retain some of the gains from capacity pooling, each ICU could be cross trained to care for the other's least sick patients, a form of "chaining". Another approach might be to search for ways to minimize the effects of variable demand. For instance, if trauma cases tend to occur on the weekends, rescheduling elective cardiac cases from Friday to Monday could create capacity when it is most needed.

CONCLUSION

Operations Management helps with the professional management of business processes. From traditional manufacturing to distribution and services, the principles and insights from Operations Management have been used successfully to help firms better manage their businesses. Determining the appropriate level of capacity is often challenging, particularly when dealing with variability from multiple sources. Operations Management provides us with the tools to determine the optimal level of capacity and to manage the trade-offs inherent in demand-supply mismatches. But Operations Management is not just about optimizing a given process or capacity allocation decision—it is also about improving process through innovation. The three examples discussed above offer a glimpse into the kinds of process innovations used by highly successful firms, but there are many more such innovations being used by firms both large and small [11]. Perhaps the greatest role Operations Management can play in the ICU is in teaching us how to apply these innovations, thereby improving both the quality and efficiency of critical care.

REFERENCES

1. Green L: **Capacity planning and management in hospitals.** *Operations Research and Health Care* 2005, **70:**15-41.
2. Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP: **Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit.** *Crit Care Med* 2007, **35:**1477-1483.
3. Lott JP, Iwashyna TJ, Christie JD, Asch DA, Kramer AA, Kahn JM: **Critical illness outcomes in specialty versus general intensive care units.** *Am J Respir Crit Care Med* 2009, **179:**676-683.
4. Likert J: *The Toyota way.* New York: McGraw-Hill; 2004.
5. Womack JP, Jones DT, Roos D: *The machine that changed the world.* New York: Simon & Schuster; 2007.
6. Jordon WC, Graves SC: **Principles on the benefits of manufactuing process flexibility.** *Management Science* 1995, **41:**577-594.
7. Kahn JM, Rubenfeld GD, Rohrbach J, Fuchs BD: **Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients.** *Med Care* 2008, **46:**1226-1233.
8. Green LV: **How many hospital beds?** *Inquiry* 2002, **39:**400-412.
9. Bazzoli GJ, Brewster LR, May JH, Kuo S: **The transition from excess capacity to strained capacity in U.S. hospitals.** *Milbank Q* 2006, **84:**273-304.
10. Girard TD, Ely EW: **Protocol-driven ventilator weaning: reviewing the evidence.** *Clin Chest Med* 2008, **29:**241-252, v.
11. Cachon GP, Terwiesch C: *Matching supply with demand: an introduction to operations management.* New York: McGraw-Hill; 2006.

LIST OF ABBREVIATIONS

ICU = intensive care unit

TPS = Toyota Production System

COMPETING INTERESTS

None

FUNDING

None

FIGURES

Figure 1. Waiting time example. In this example a sample process takes an average of four minutes and 12 customers arrive randomly per hour. Time (in minutes) is on the Y axis. The top figure shows total process time, with service time is shown in blue and wait time is shown in red. The bottom figure shows the number of customers in the process at any one time.

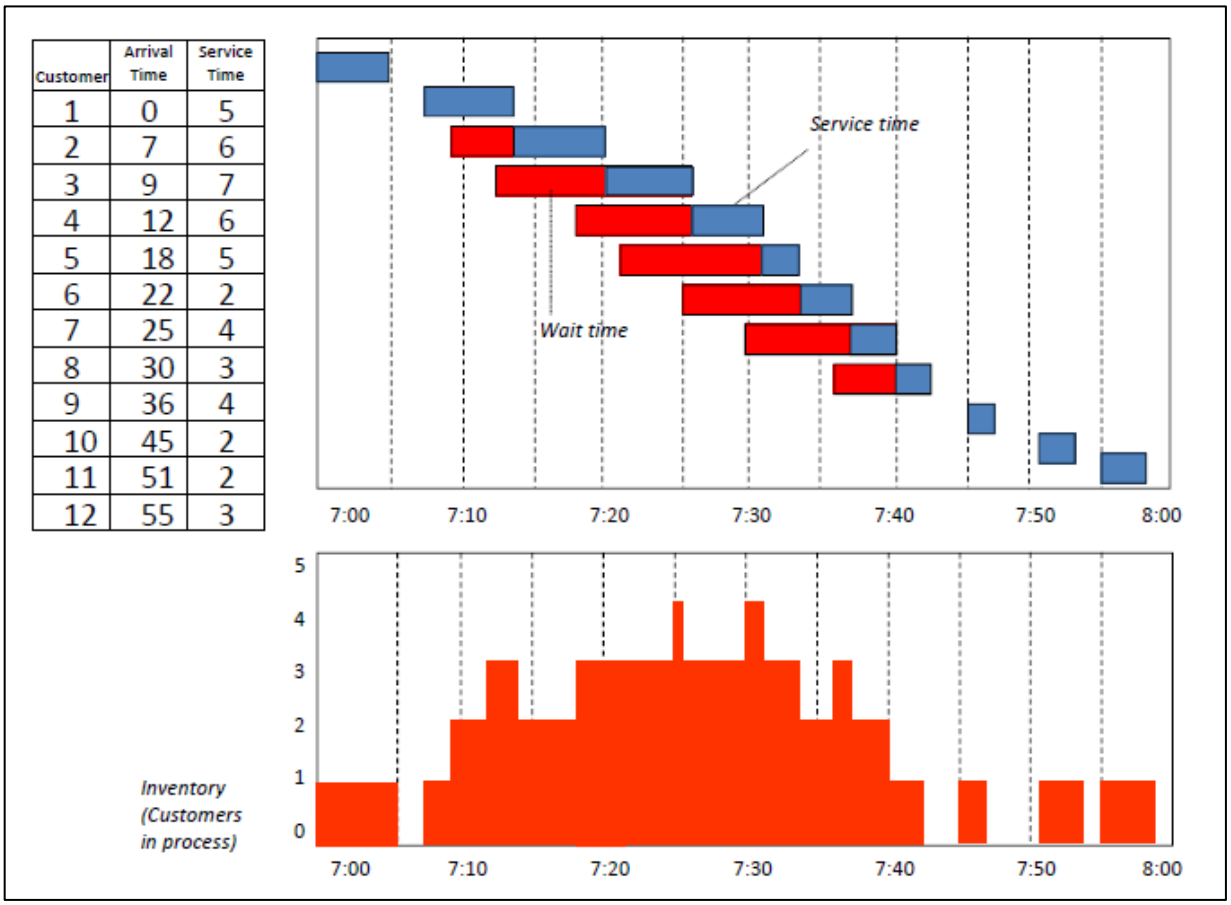
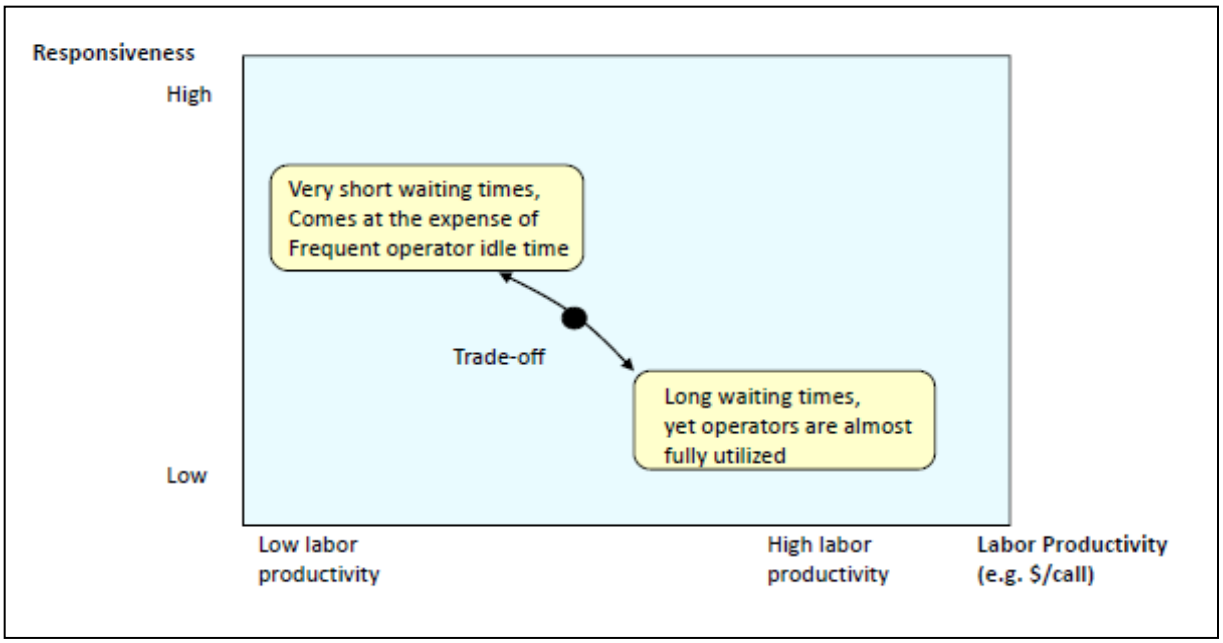


Figure 2. The interaction of process responsiveness and productivity. The top panel shows the tradeoff between responsiveness and productivity in a given process. The bottom panel shows how process redesign can improve both responsiveness and utilization.



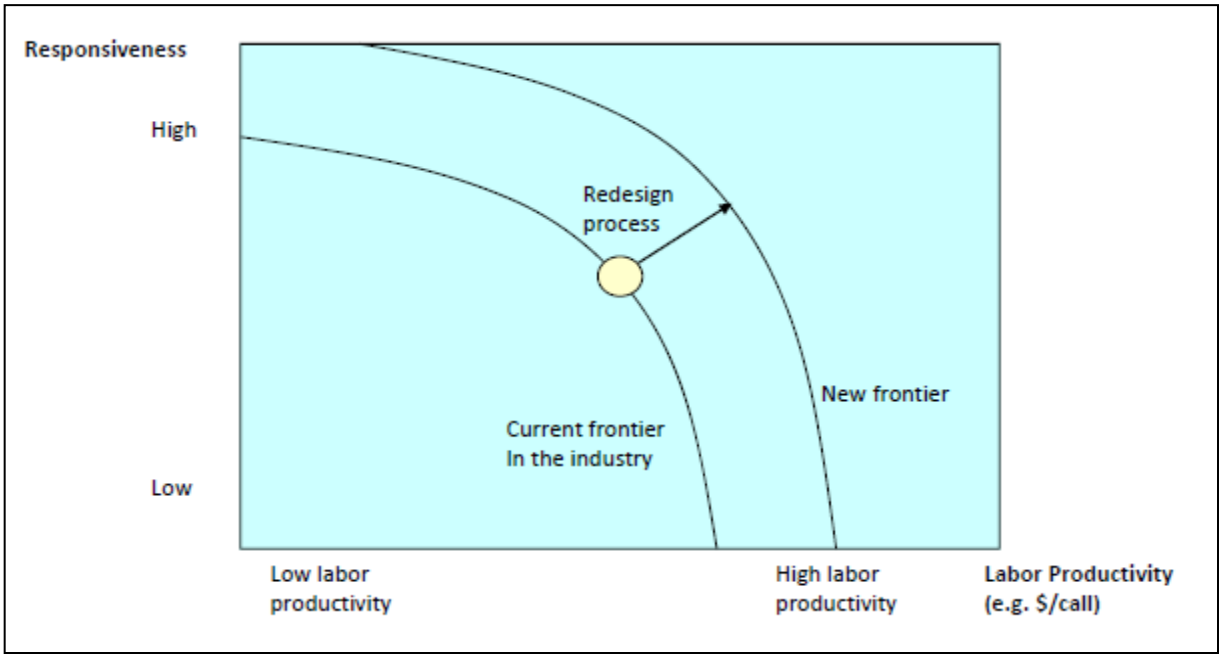


Figure 3. Productivity comparison in the US airline industry. Compared to other US airlines, Southwest achieves similar yields with greater efficiency. Lufthansa and Ryanair are added as non-US illustrative benchmarks. ASM=available seat mile; RPM=revenue passenger mile.

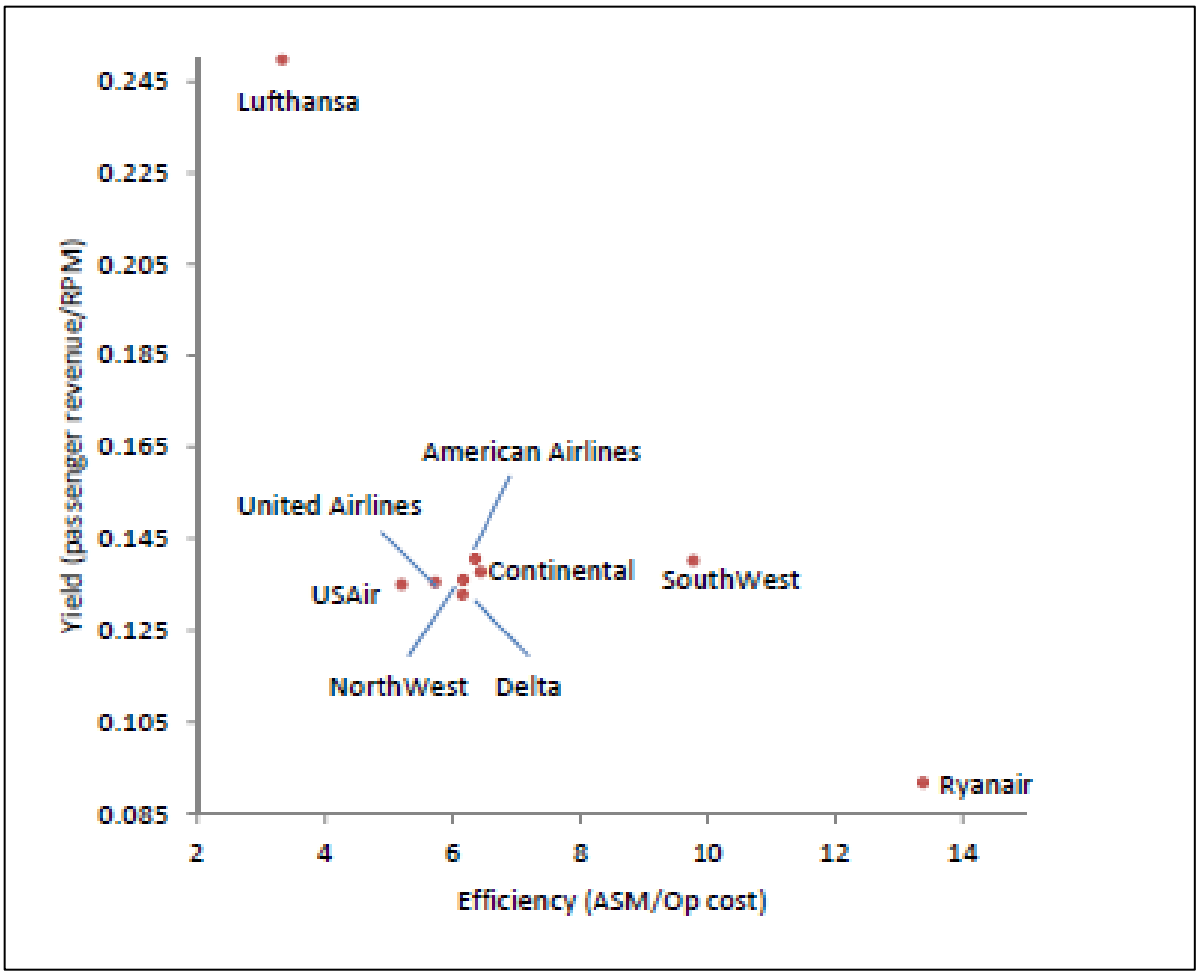


FIGURE 4. Partial versus full flexibility. Partial flexibility can be almost as beneficial as full flexibility, especially if the production system is designed in a way that every product can be produced at two plants and the assignments form a "chain" as illustrated on the left.