

# **DEFINING LEAN ACCOUNTING: UNDERSTANDING THE IMPACT OF COST ACCOUNTING SYSTEM DESIGN ON LEAN MANAGEMENT**

Robert Hutchinson, Oakland University  
Phone: 248.370.4283, Email: [hutchin2@oakland.edu](mailto:hutchin2@oakland.edu)

## **INTRODUCTION**

Despite a plethora of literature on the subject, a great deal of confusion continues to surround the concept of lean accounting. One key source of confusion is the fact that the concept of lean accounting is expressed along two different dimensions. The first dimension refers to what extent a management accounting system supports lean management principles in the long-run, and the second dimension refers to the efficiency of the accounting system itself in delivering needed information to management in a timely and cost effective manner.

Many managers have become increasingly aware of the inherent problems with their legacy cost accounting systems, particularly along the first dimension. Traditional costing systems often create perverse incentives for managers that undermine the principles of lean management. Practitioners have looked to management accounting researchers and consultants to offer alternatives that would be lean along both dimensions. Given the attention activity-based costing (ABC) and more recently resource consumption accounting (RCA) have received in the management accounting literature, it appears that a chasm between operations management and management accounting theory has grown wider. A recently published paper asks the question “Are ABC and RCA Accounting Systems Compatible with Lean Management?” (Grasso 2005), but misses a key nuance of capacity management theory that puts capacity accounting directly at odds with lean management.

According to the author, “the cost of developing and maintaining an RCA system far exceeds the benefits for lean businesses. It would be hard to imagine a lean-oriented company adopting RCA. From an accounting perspective, it would also be hard to characterize a company using an RCA system as lean”. While this statement is certainly true, it focuses more on the cumbersome nature of RCA and misses the far more important any derivation of capacity accounting is absolutely incompatible with lean management. The author’s statement “excess resources (i.e. unused capacity) are a waste” while widely accepted as common sense, neglects capacity dynamics in the context of a lean organization.

This paper looks at the historical roots of capacity accounting and lean management, to better understand how they evolved and in what competitive environment they would make sense, and builds on the previously mentioned paper by examining the nuances of capacity management, and demonstrate how capacity accounting undermines the very essence of lean management.

## **THE TEUTONIC ORIGINS OF CAPACITY ACCOUNTING**

Resource Consumption Accounting, and any other form of capacity accounting, has received a great deal of attention recently in the management accounting literature. In a one sentence description, capacity accounting uses the total amount of the allocation base at capacity to calculate overhead rates and assign cost to cost objects (Garrison et al. 2008). The idea of using the theoretical capacity as a basis for allocating manufacturing overheads is not a new one; rather it is derived from German cost accounting practice that has long focused on capacity utilization. The use of such capacity accounting methods has historical roots dating back to the very industrialization of Germany.

Germany, which did not have colonies in South America, Asia, and Africa, was at a major disadvantage when compared to the other European powers with their ready access to the raw materials demanded by industrialization. In order to compete with the burgeoning textile industry in England, Germany was forced to create its own dyes synthetically. Hence, Germany became an early leader in the chemicals sector. Early on, chemicals production required relatively higher investments in fixed capital and lower levels of variable manufacturing costs than the textile industry it served. With such a high degree of operating leverage, it therefore makes sense that cost accounting systems that identify and quantify underutilized production capacity would have Teutonic origins.

Today’s proponents of capacity accounting say that it addresses two main concerns of traditional costing systems that derive their predetermined overhead rates based on estimated or budgeted activity in the coming period. First, since the predetermined overhead rate is based on “actual” capacity, it will not fluctuate from period to period with budgeted activity. Second, products are only charged for the portion of the resources they actually use.

Any sort of capacity-based accounting system virtually ensures that there will be underapplied overhead at the end of any accounting period. This amount of unallocated overhead can be treated in two ways, either allocated between Cost of Goods Sold and Finished Goods inventories or treated as a period expense. Proponents argue, however, that in order for capacity accounting to have a real impact underapplied overhead should not be buried in inventory accounts. They suggest that underapplied overhead really represents the Cost of Unused Capacity, which should be reported as a period expense. To illustrate this point, assume the following production data:

Actual volume	40,000	units
Selling price	\$40.00	per unit
Variable production cost	\$24.00	per unit
Fixed MOH	\$100,000	per year
Capacity	50,000	units
Fixed SG&A	\$500,000	per year

Under the traditional approach to overhead allocation, the predetermined overhead allocation rate would be \$2.50 per unit (\$100,000 / 40,000 units budgeted volume). Under the capacity accounting method, the allocation rate would be \$2.00 per unit (\$100,000 / 50,000 units at capacity). The income statements appear as follows:

Traditional Method		Capacity Method	
Revenue	\$ 1,600,000	Revenue	\$ 1,600,000
Cost of goods sold	<u>1,060,000</u>	Cost of goods sold	<u>1,040,000</u>
Gross margin	540,000	Gross margin	560,000
SG&A expense	<u>500,000</u>	<b>Cost of idle capacity</b>	<b>20,000</b>
Net operating income	<u>\$ 40,000</u>	SG&A expense	<u>500,000</u>
		Net operating income	<u>\$ 40,000</u>

Given the old adage, “you get what you measure”, no doubt that factory managers will then find ways to ensure that this period expense is minimized. Understanding the inverse relationship between capacity utilization and the cost of idle capacity, the managerial incentives created come into direct conflict with the principles of lean management.

## **THE JAPANESE ORIGINS OF LEAN MANAGEMENT**

The concept of lean management has very different origins to capacity accounting. The practices which we call lean management were originally laid out in Taiichi Ohno's (1988) seminal work, *Toyota Production System*. Ohno had worked his way up through the ranks of Toyota, first starting out in the Toyoda family's loom business, and prior to the war, moving into the startup Toyota Motor Company. The Toyota Production System was born out of necessity, as post-War Japanese economic growth demanded a greater variety of goods than could be economically mass produced.

Since lean management was first introduced in the United States in the early 1980s, a great deal of literature has been written on the subject. Unfortunately, the popularity of the subject itself may add to some of the confusion. Because the origins of lean management are in the Toyota Production System, referring back to the original work may clear up a lot of misconceptions, particularly when it comes to the issue of capacity utilization. This paper will make frequent references to Ohno's original work, first published in Japanese in 1978. It will then demonstrate from an operations management perspective that not only is RCA incompatible with lean management in the sense that it itself is not lean accounting process, but that any form of capacity accounting actually serves to undermine the very principles of lean management laid out by Ohno in Toyota Production System.

### **TOYOTA PRODUCTION SYSTEM'S VIEW ON CAPACITY UTILIZATION**

At its essence, Toyota Production System seeks to eliminate waste, in all its forms, from the production process. This brings us to a key sources of confusion vis-à-vis capacity management. To bring this into better focus, it is necessary to first realize that kaizen or continuous improvement is a journey and not a destination. Lean management understands that waste will never be totally eliminated; therefore, it is necessary to prioritize which types of waste are most inefficient.

One of the primary tenants of TPS is the concept of just-in-time (JIT) production, which is to say that production is pulled by actual customer demand. JIT therefore seeks to reduce and if possible eliminate inventory. Inventory is by far one of the most nefarious of all forms of waste, because it not only ties up working capital, but it actually hides production inefficiencies and defects. Despite our understanding of the wastefulness of maintaining inventory, the building of inventory seems to be culturally engrained. According to Ohno:

Accepted wisdom tells us if a new machine is purchased, keep it operating full-time... As long as it is running smoothly, let the machine produce to capacity... In case of future trouble with the machine, let it produce while it can. This way of thinking is still deeply rooted among manufacturing people (pg. 101).

With volatility in customer demand, simply eliminating inventory can cause serious problems with service level. This is where excess production capacity becomes a necessary aspect of TPS. Put quite simply, TPS trades inventory buffers for capacity buffers and actually advocates maintaining excess productive capacity whenever possible. As Ohno had written:

Let's consider Toyota's thinking about what is economically advantageous from the standpoint of production capacity. Opinions differ on the economic advantages of maintaining extra production capacity. In brief, excess capacity utilizes workers and machines that are otherwise idle, incurring no new expense. In other words, they cost nothing (pg. 56).

The principle of excess productive capacity simplifies managerial decision making under several different scenarios. In each case, knowing that excess productive capacity exists virtually ensures that

managers will make the correct decisions for long-term competitive advantage under the following scenarios:

- Make or buy decisions become a marginal (variable) analysis, where only the additional cost of materials and labor are relevant.
- Preventive maintenance and line work does not require any cost consideration; there is no marginal cost.
- Reducing lot sizes carries no marginal cost, and reducing setup times, therefore, becomes a separate issue not immediately affecting production.

Again, in the words of Ohno:

When there is excess capacity, loss or gain is evident without requiring cost studies. The most important thing to know is the extent of excess capacity at all times... At Toyota, we go one step further and try to extract improvements from excess capacity. This is because, with greater productive capacity, we don't need to fear new cost (pg. 57).

Ohno makes a key distinction between the operating (rated capacity) versus operable rate (scheduled capacity) of a machine process, placing a greater emphasis on the latter. Rated capacity, sometimes called engineering capacity, is the maximum theoretical speed at which a machine can operate assuming no breakdowns, failures, or shutdowns. Even if achievable, it can only be sustained for brief periods. Scheduled capacity is the expected standard rate or speed. Actual capacity includes downtime for breakdowns, stoppages, and regular maintenance, as well as allowances for yield problems. What is most important for TPS is that production capacity is always available when it is needed, which is what Ono refers to as 100% of operable rate:

The operating rate is the current production level in relation to the full operating capacity of the machine for a specified length of time. If sales go down, the operating rate naturally drops. On the other hand, if orders increase, the operating rate can reach 120 percent or more through shift work and overtime. Whether an operating rate is good or bad is determined by the way equipment is used relative to the quantity of products needed. The operable rate at Toyota means a machine's availability and operable condition when the operation is desired. The ideal 100 percent depends on good equipment maintenance and rapid changeovers (pg. 126).

Capacity is highly dependent on the product mix being produced. Capacity can be stored, i.e. buffer capacity. Capacity is dynamic and tends to expand overtime even without major new investment through the "learning" effect. Capacity is highly affected by the degree of variability of demand and processing time. Idle time results if a process completes its work unusually quickly and no other jobs are waiting to be worked on.

### **CAPACITY MANAGEMENT IN A LEAN ENVIRONMENT**

Let us look at a brief quantitative example to demonstrate the impact of increased capacity utilization on manufacturing system performance, and in particular, on the levels of work-in-process (WIP) inventory. Assume that we are talking about a lean manufacturing system with JIT production; hence production only takes place in the presence of actual customer demand. Now suppose a workstation has a engineering (operating) capacity of 20 pieces per hour.

Suppose that both the arrival time ( $a$ ) of work orders and the time it takes to process the orders ( $p$ ) are governed by exponential probability distributions. This is a reasonable assumption if we assume that customer orders arrive one at a time and the probability of arrival during a small time interval of

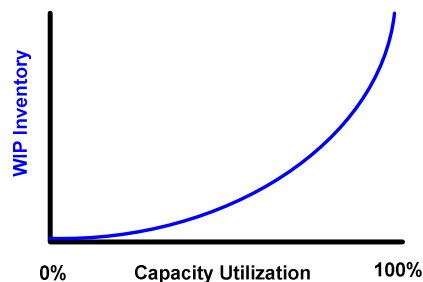
constant duration does not change. That is to say, the probability of a customer order arriving does not depend on the time of day, is independent of previous customer orders, and independent of the number of customer orders currently in queue. Analogously, this implies that the probability of the workstation completing any job in the given small amount of time remains the same, no matter how many jobs are waiting in queue or how much time an order has already spent in process.

Now, what happens if customer orders arrive at an average rate of 12 per hour, the probability that one will arrive during any given second during that hour is  $1/300^{\text{th}}$  or .00333 (=12/60/60). In general, if the probability of a customer order during a short period of time  $\Delta t$  is equal to  $(a \times \Delta t)$ , then the average time between arrivals is  $1/a$ , therefore the interarrival rate is  $1/12^{\text{th}}$  of an hour, or  $60/12 = 5$  minutes.

Given these assumption of an exponential arrival and processing rates, it can be shown that if the arrival rate is  $a$  (12 in our example) then the probability of no customer orders arriving during a long period of time from  $(T)$  is equal to  $e^{-aT}$ . Similarly, the processing rate  $(p)$  would have an average processing time of  $1/p$ , therefore the percentage of time the workstation is idle during any time period is equal to  $(1 - a/p)$ . Therefore, if orders arrive at a rate of 12 per hour and the machine can process them at a rate of 20 per hour, it will be idle about 40 percent of the time. Now comes the important point, that even though the machine process has 40 percent excess capacity, we might expect to see WIP inventory because some of this capacity is lost to enforced idle time. In this case the average number of orders either being processed or in queue  $N$  can be calculated as  $N = a / (p-a)$  or 1.5. So even with 40 percent excess capacity, we would expect to see an average of .5 orders in queue.

In the case of exponential interarrival and service times, the average time an order spends in a stable system is from the time it arrives until it is completed si equal to  $T = 1/(p-a)$ . This is because the average number of people in a line is equal to the average rate at which they arrive times the average amount of time each order spends in queue or  $N = a \times T$ . This implies that  $T = N/a$  or  $a/(p-a) \times 1/a$  or  $1/(p-a)$ . This is based on the famous “Little’s Law” named for the man who first proved that critical WIP is equal to arrival time times cycle time. To calculate the total average cycle-time per order, we simply multiply the number of orders in process or in queue  $(N)$  times the average interarrival time  $(1/a)$  of .125 hours or 7.5 minutes.

What is most important from this example is the impact on WIP inventory as capacity utilization increases. For example, suppose that the order arrival rate increases by 50 percent to 18 per hour, still well within the rated capacity of our workstation. In this situation, the WIP inventory  $(N)$  would increase 600 percent from 1.5 to 9 and the average cycle time would have a corresponding increase from 7.5 minutes to 45 minutes! What lean manufacturers understand is that in the face of demand uncertainty, as capacity utilization approaches its maximum, overall manufacturing system performance deteriorates rapidly. The general relationship between capacity utilization and WIP inventory is as follows:



One other nuance of capacity management that Ohno understood well was the difference between making improvements to existing machine processes versus adding new machines as a strategy for creating capacity buffers. Returning to our previous example, two decisions management might make in the face of the increasing demand as customer orders arrive at 18 per hour and WIP inventory approaches 9, are to either add another machine or make improvements to the existing machine in order to increase its processing rate.

Looking at the first option, let us suppose that an additional work station is added. In this case, the arrival rate for each individual machine station would be cut in half to 9 per hour. We would therefore expect the number of customers in each queue would fall to  $.82 [=9 / (20 - 9)]$  so that the number of orders in both lines would be 1.64 and the average time an order would wait in the system would be 5 minutes and 28 seconds ( $.82 \times 1/9 \times 60$  minutes).

Now, instead suppose that an improvement could be made to the existing machine that would cut its processing time in half so that it can produce 40 units per hour. Then, according to the equation the number of customers would fall to  $.82 [18 / (40 - 18)]$ , but the average waiting time for an order would be only 2 minutes 44 seconds ( $.82 \times 1/18 \times 60$  minutes) or half the time under the two machine process approach! In other words, doubling the capacity rate of an existing slower process is more effective at lowering inventory than doubling the number of slower processes. This fact has major implications for management accounting practice.

#### **IMPLICATIONS ON MANAGEMENT ACCOUNTING PRACTICE**

Capacity accounting, much like ABC, fails to add value vis-à-vis Japanese management techniques. While ABC is certainly a failure along the first dimension of a lean accounting system, i.e. that it is in itself not a lean process, capacity accounting is also difficult to classify as a lean process. Simply trying to calculate productive capacity is near impossible in most flexible manufacturing environments. Even for a single facility there usually is considerable uncertainty as to how its capacity ought to be measured for planning purposes. For example, an operation's rated capacity is different from both its scheduled capacity and actual capacity (Hayes et al. 2005).

Capacity accounting actually undermines core principles of lean management in that it creates perverse incentives for managers to overproduce and postpone preventive maintenance in an effort to increase capacity utilization of key equipment. If anything, unused production capacity can be viewed as a competitive advantage. The more common view of capacity cushion is that associated with resources such as floor space, equipment, and people. This kind of capacity cannot provide the same speed of response as can inventories, but it is more flexible, in that the specific mix and volumes of products demanded by customers can be produced within the company's normal lead time. In a market that is growing faster than expected, it might enable you to attract new customers who are not getting the service they desire from your competitors who are short of capacity. It also might allow you to take market share from competitors who are concerned more about their near-term profitability and return-on-investment than their long-term market position.

**\*\*\*Please contact author for list of references.**