CHAPTER 1

Understanding the Fundamentals of the Theory of Constraints

OBJECTIVES

• Define precise operational measurements for evaluating productivity under TOC
• Define productivity measurements in the context of a healthcare system
• Explain bottlenecks
• Introduce the five focusing steps for the process of ongoing improvement

When I (Chris) was in the fifth grade, our teacher gave us a six-page test and told us it was timed. As soon as she said go, I tore through the pages, determined to finish first. The test consisted of a series of questions and tasks: “Draw three triangles inside three circles.” “When red and blue are mixed together, what color do they make?” The questions weren’t difficult. Halfway through the test, the instructions said to speak your name out loud. I was the first in the class to reach this point. Three pages later, I was a few items away from finishing. I was sure I was ahead of everyone else. I looked up at my best friend Brad, who was sitting at his desk—hands folded in front of him, pencil off to the side. What the heck? But I was too close to being first to worry about Brad. Toward the bottom of the last page, I read: “Say ‘I’m almost done’ out loud.” Again, I was the first in the class to do so. The second-to-last item read: “Don’t complete any items except for the first and the last.” The final item: “Sit quietly with your hands folded in front of you and your pencil off to the side.” I nervously went back to item number one, which I now realized said: “Read all items first before starting.” Forty years later, I can still see Brad looking at me and smiling, and I can still feel my face turning red.
My takeaway from that experience was: Next time, cheat off Brad. Another, more enduring takeaway is that it is important to think through any strategy completely and critically before jumping in. Doing so will save a lot of time (and maybe some embarrassment). It would be great to know if an idea that was meant to improve patient flow actually would improve patient flow before committing time and resources to it. You have probably worked on a committee that successfully executed a project everyone assumed would help improve flow, only to find that after countless meetings and hours of implementation, the project didn't move the needle. A project can be completed without being effective. In *The Goal*, Eliyahu Goldratt refers to this as a “nonproductive move.” It is important to determine whether a project will be productive or nonproductive prior to getting started.

TOC uses three measurements to predict whether a project will be productive or nonproductive: throughput, inventory, and operating expense (exhibit 1.1). While these measurements may seem simple and intuitive, their definitions are precise:

- **Throughput** is the rate at which the system generates money through sales.
- **Inventory** is all the money the system invests in purchasing things that it intends to sell.
- **Operating expense** is all the money the system spends to turn inventory into throughput.

These measurements can be applied to virtually any system where materials move through a stepwise process to a finished product—for example, a widget factory that starts with raw materials, alters these raw materials through a series of steps,
and ultimately makes a finished widget. The materials and product can be broadly defined and are relevant to systems outside of the manufacturing sector.¹

For a move to be productive, it should increase throughput while simultaneously decreasing inventory and operating expense (exhibit 1.2). At the very least, any move must affect two of these measurements while holding the third static.

Consider, for example, a Mattel factory that manufactures Barbie dolls. Inside the factory, raw materials are combined and processed through a series of steps to create a finished product. For simplicity, assume that each Barbie doll is made by combining six parts: head, body, two arms, and two legs. These individual body parts constitute the raw materials.

The throughput is the rate at which the individual body parts (the raw materials) are combined into a completed Barbie doll and the doll is packaged and sold for money. The inventory is the money the company spends to purchase body parts. Inventory also includes the finished dolls that have not yet been sold as well as partially completed dolls that are somewhere on the assembly line (also known as works-in-process). Operating expense is the money used to turn the body parts into the completed Barbie doll as well as the costs associated with storing, selling, and shipping the completed product and any other costs associated with turning raw materials into finished dolls.

Any business decision Mattel makes regarding Barbie doll production is potentially disruptive and costly, but it may also reap huge rewards for the company. The managers at Mattel can use operational measurements to predict whether a business decision is likely to be productive or nonproductive. For any change in operations to be productive (i.e., make more money), that change must simultaneously result in the following outcomes:

Exhibit 1.2: Effect of a Productive Move on Operational Measurements

![Diagram showing the effect of a productive move on operational measurements: throughput increases, inventory decreases, and operating expense decreases.]

¹This is an unedited proof.
Copying and distribution of this PDF is prohibited without written permission.
For permission, please contact Copyright Clearance Center at www.copyright.com
• Making and selling Barbie dolls more quickly (increase throughput)
• Purchasing fewer body parts at a given time and storing fewer works-in-process and completed products while still meeting demand (decrease inventory)
• Reducing overall costs to the organization (decrease operating expense)

Note that making Barbie dolls more quickly does not increase throughput unless Mattel also sells the dolls at a faster rate. Speeding up the manufacturing process in isolation may result in increased inventory and operating expense, which is the opposite of the desired goal.

It bears repeating that for any action to be productive, it must simultaneously result in an increase in throughput, a decrease in inventory, and a decrease in operating expense. However, of the three variables, throughput is the most important. If all you do is increase throughput, the entire system still benefits. Customers receive more products and services, the company improves its finances, and employees have more stability. Throughput is so important that an action can be productive even if it increases operating expense, as long as it increases throughput by a higher amount.

OPERATIONAL MEASUREMENTS IN HEALTHCARE

How do these operational measurements apply to patients in a hospital setting? If you consider each patient as analogous to a product, you can see how TOC is used in patient care. In a factory, a product starts with parts or materials that go through a stepwise process to be altered or assembled into a finished product, which is then sold. Throughput measures how quickly this process occurs. Inventory is the money tied up in raw materials, works-in-process, and unsold finished product. Operating expense is the money it costs to turn inventory into finished product (including the cost of storing the parts and unfinished products until they are sold).

In healthcare, a patient starts with a symptom, abnormal test result, or health concern and must go through a stepwise process to be diagnosed, treated, and discharged. In the healthcare setting, the measurements of throughput, inventory, and operating expense apply directly to patient flow through a hospital, clinic, or office-based practice in the following context:

• **Throughput** is the rate at which a patient moves through a hospital, clinic, or office-based practice.\(^2\)
• **Inventory** is the number of patients at different stages in the hospital, clinic, or office-based practice.
• Operating expense is all the resources used in evaluating, diagnosing, treating, and discharging patients, plus all the other costs associated with running the hospital, clinic, or office-based practice.

Throughput can be measured for an entire hospital or for one specific area or unit. For example, ED throughput is measured from the time a patient registers in the ED to the time the patient is discharged, admitted and transferred to an inpatient unit, or transferred out of the ED to another hospital. Inpatient throughput is measured from the time a patient arrives at a unit (e.g., medical-surgical, telemetry, critical care) to the time the patient leaves the unit (e.g., transfers out of the intensive care unit [ICU], is discharged from the hospital ward). Throughput can also apply to a patient’s entire hospital length of stay, from the time the patient first registers to the time the patient has physically left the hospital. To improve throughput, you must change the process for evaluating, diagnosing, treating, and discharging patients.

Since throughput is the rate at which the hospital generates money through sales, it must include the entire time a patient is physically in a hospital bed, even if the patient’s discharge order has been written (patients are considered inventory until they no longer physically occupy their hospital beds, just as a widget is considered inventory until it is actually sold and moved out of the factory). Examples include an ED patient who is being admitted but is boarding in the ED or an inpatient who has been discharged but is still in the hospital room waiting for a ride home. Patients boarding in the ED are one of the biggest indicators of poor flow in a hospital and are often a starting point for identifying bottlenecks.

Similar to throughput, inventory can be measured by patient census over the entire hospital or in a specific department or unit. Examples of inventory include the number of patients in the ED, ICU, or operating suites. Inventory can also be defined as a subset of patients in a department, such as patients in the ED waiting room who have yet to be evaluated or admitted patients who are boarding in the ED. Inventory can even consist of patients awaiting treatment who are not in the hospital but may be waiting at home for a procedure or clinic appointment.

Operating expense refers to the money spent to run the hospital: salaries for physicians, nurses, technicians, and housekeepers; utility bills; hospital supplies; food; and linen costs. Essentially, operating expense includes all the money that the hospital spends to treat patients. Operating expense also includes money that the hospital spends independent of patient census: upkeep on imaging equipment, for example, or licensing fees. The hospital incurs these expenses even if it has no patients to treat.
As in manufacturing, the goal in healthcare flow is to increase throughput, decrease inventory, and decrease operating expense. For a hospital, this means faster dispositions (the rate at which patients are evaluated, treated, and discharged), decreased time that a patient stays in the hospital (shorter lengths of stay, fewer avoidable days), and more efficient treatment plans (judicious use of laboratory and imaging studies, standardized supply chain, implementation of evidence-based protocols).

We can illustrate the power of TOC operational measurements by applying its basic principles to a case study (adapted from Han et al. 2007).

**CASE STUDY**

An urban, academic Level 1 trauma center with 45,000 annual visits underwent a major expansion to increase the size of its ED from 28 to 53 licensed beds. The expectation was that the expansion would cause ambulance diversion times to decrease, but when the hospital studied the effects of the expansion on ED flow, it found no significant change in ambulance diversion times, the number of episodes of ambulance diversion, or the duration of each episode. Moreover, both the total ED length of stay and the length of stay for boarded patients in the ED increased after the expansion.

What happened?

Let’s answer the question of what happened in terms of our three operational measurements: throughput, inventory, and operating expense. To be a productive move, doubling the number of ED beds would need to increase throughput, decrease inventory, and decrease operating expense. What effect did the expansion have on each measurement?

The hospital believed it could accommodate more emergency patients by doubling its number of ED beds. Does this move improve throughput? Remember that throughput is the rate at which a patient moves through a department, not just into a department. The hospital made no changes to its process of evaluating, treating, or discharging patients and, therefore, has not improved throughput. Some patients will now wait to be evaluated, treated, and discharged in a bed instead of the waiting room, but whatever factors previously limited patient flow are still in effect. The new beds will soon be filled with additional patients who are parked in the ED. Doubling the number of ED beds makes the parking lot bigger, but it doesn’t move cars out of the lot faster.

What effect does the expansion have on inventory? If the rate at which patients leave the ED is unchanged, then inventory will probably not go down.
The number of patients in the waiting room may decrease initially, but an equivalent number of patients will be parked in the newly built ED beds. The ED has become a bigger parking lot, and inventory will soon increase.

How does this move affect operating expense? Operating expense will increase; the additional beds will require more staff to maintain nursing ratios, more work for housekeeping, more supplies to stock the new rooms, and more utility costs. A bigger parking lot is more expensive.

Building more beds in the ED without improving the process of diagnosing, treating, and discharging patients either will not affect throughput or will decrease it somewhat (due to added inefficiencies that result from boarding more patients). An expansion will increase inventory and operating expense. It is a nonproductive move.

Hopefully, it now seems obvious that the plan in the case study was doomed to fail. After reading only a few pages, you have become more skilled in patient flow than the hospital leaders who approved the ED expansion. TOC has provided insight that could have saved tens of millions of dollars on an expansion that did not achieve the hospital’s goals. You have learned that before starting a project, you must first ask if the project will be productive, and you have started to learn how to answer that question. Operational measurements seem intuitive and simple to understand, but they are both powerful and often ignored.

To summarize:

- Any move that increases throughput, decreases inventory, and decreases operating expense is productive, and any move away from this goal is nonproductive.
- Throughput is the most important of the three measurements.
- The operational measurements must be applied to all potential projects before beginning any project.
- A project can be successfully executed, but that doesn’t mean it will be productive.

Our objective, then, is to find actions that are productive.

UNDERSTAND THE BOTTLENECKS

A bottleneck is any resource whose capacity is equal to or less than the demand placed on it. Every system has (at least) one main bottleneck that limits the rate of output for the entire system. It is often the most critical resource of the process, and its capacity may not be easily increased. Bottlenecks can occur due to a temporary situation, such as an unexpected increase in demand, a machine breaking down,
Part I: Smashing the Bottleneck: Moving Patients Through the Hospital

Bottlenecks may also be self-inflicted, the result of practices and policies such as batching work (e.g., doing all elective surgeries on Mondays) or excessive multitasking.

The key to TOC is to identify the scarcest resource in a supply or service chain and improve the use of that resource, thereby improving flow through the entire system. Let’s create a hypothetical scenario for our Barbie doll example to illustrate a bottleneck. Assume that Mattel needs a million pieces of each of the six body parts to meet its production target. It has one million pieces for five of the parts, but only one thousand pieces of the right arm. In this example, the right arm is the bottleneck.

A nonbottleneck is any resource whose capacity is greater than the demand placed on it. All of the body parts other than the right arm are nonbottleneck resources. Since there is a shortage of right arms, there are more of the other body parts than the factory can use at the present time.

A bottleneck resource’s productivity determines the productivity of the entire system. Bottlenecks are important because Mattel can produce only as many Barbie dolls in an hour as there are right arms available. If the factory has the capacity to produce one million dolls in an hour but has only one thousand right arms (the bottleneck) available in that hour, then the entire factory can make only one thousand dolls.

An hour lost on the bottleneck is an hour lost for the entire system. Once the bottleneck resource is used up (no more right arms), the factory sits idle. It can’t produce any more finished products without additional right arms.

The cost of an hour lost on the bottleneck is the cost of an hour for the entire system. While the factory sits idle, waiting for more right arms to be delivered, it is not producing finished products and, therefore, is not generating revenue. However, the factory’s fixed costs do not change, and it continues to incur expenses roughly equal to the expenses it incurs when it is producing goods to sell.

An hour saved on a nonbottleneck is a mirage that does not affect productivity. Let’s say that at the Mattel factory, there are six teams of workers and that each team is responsible for one of the body parts. The factory manager may notice the right-arm team sitting idle while the factory waits for a shipment of right arms. The manager transfers the right-arm team members to the other teams, believing that the plant’s efficiency can be improved by eliminating downtime. Now each of the five other teams is faster at assembling its respective part—but this action has not saved the factory any time because the entire factory is waiting on the availability of right arms to complete the product. This move has allowed the factory to make one-armed Barbie dolls more quickly, but since Mattel can’t sell one-armed Barbies, throughput (the rate at which Mattel makes money through Barbie sales)
is not affected, and time is not actually saved. Moreover, the manager’s decision has driven up inventory (as there are now 999,000 one-armed Barbies piled up behind the right-arm station) and increased operating expense (since Mattel now has to pay to store the unfinished products). Mattel did not make additional money despite improving the speed of the nonbottleneck teams by keeping all the workers busy.

Saving time on a nonbottleneck introduces the concept of local optima. In a system where departments operate as silos, a change made in one department may seem to improve efficiency but only if productivity is measured for that specific department. When you look at the overall picture, the change may prove nonproductive (e.g., making 999,000 one-armed Barbies). Further, local optimization may reduce throughput through the entire system, increase inventories and operating expense, and move the system away from productivity. If saving time on nonbottlenecks increases inventory, increases operating expense, and makes the plant more chaotic (by creating more works-in-process and piles of inventory), the organization has moved away from its goal.

**FIVE FOCUSING STEPS**

Process improvement requires identifying a system’s bottleneck resource and then breaking it. In *The Goal*, Goldratt outlines five steps, called focusing steps, used to identify and break the bottleneck. Note that every system has a bottleneck; bottlenecks are neither good nor bad—they are simply an inherent part of any process. Once one bottleneck has been broken, another will appear. For this reason, TOC is a process of ongoing improvement.

The 5FS that comprise the foundation of TOC are:

1. **Step 1**: Identify the system’s bottleneck.
2. **Step 2**: Decide how to exploit the system’s bottleneck.
3. **Step 3**: Subordinate everything else to the decision made in Step 2.
4. **Step 4**: Elevate the system’s bottleneck.
5. **Step 5**: If the bottleneck has been broken, go back to Step 1.

Step 1 is to identify the bottleneck. You do that by looking for a pileup of inventory in the system; the bottleneck will most likely be the resource in front of the highest pileup. In Step 2, you decide how to exploit the bottleneck. This means developing the necessary strategies to keep the bottleneck resource up and running every hour of every day. Remember that an hour lost on the bottleneck is an hour lost for the entire system. Step 2 of the 5FS is the heart of the
TOC operational strategy. Deciding how to exploit the system’s bottleneck will determine how the rest of the resources are managed. You execute Step 3 by deciding how your other resources will support the activities designed in Step 2. Step 3 requires you to prioritize the constant production of the bottleneck above all other activities—including subordinating personal agendas and institutional methodologies. In this step, you identify and reengineer local policies to benefit your system as a whole. This involves breaking down silos and changing culture, making it the most involved and meatiest of the five steps. Step 4 is to elevate the bottleneck: If, after exploiting the bottleneck resource, there is still need for greater capacity, you may need to invest in additional quantities of the bottleneck resource or in other resources that can augment the bottleneck. Finally, once you have broken the bottleneck, you return to Step 1 and look for the new bottleneck in your system. Without the vigilance of Step 5, the system will revert back to its original state and you will lose any gains made from the 5FS.

Let’s apply the 5FS to our Mattel factory.

Step 1: Notice the pileup of left-armed Barbie dolls. The inventory pileup sits behind the bottleneck resource—the right arm.

Step 2: Decide the right-arm assembly station should never sit idle: There should never be any down time at this station.

Step 3: Subordinate all other factory workflow to the decision made in Step 2. Reorganize the workflow so that the right arm is assembled first. The right arm assembly will never have to wait on the other body parts (the nonbottlenecks) to be assembled. As soon as another right arm becomes available, make sure it gets attached to a Barbie. When no more right arms are available, stop assembly until you obtain more right arms. This avoids pileups of works-in-process.

Step 4: Increase capacity of the bottleneck. Find other factories that can supply right arms; salvage right arms from other, unsold types of Barbie dolls (e.g., Lawyer Barbies and Rock Star Barbies); purchase extra left arms and turn the hands around; buy surplus G.I. Joe right arms and come out with a special edition Barbie with Kung-Fu Grip.

Step 5: Recognize when right arms are no longer the bottleneck, and go back to Step 1 to find the new bottleneck.

Now that we have introduced the 5FS, we can discuss how to apply them to healthcare.
SUMMARY

• TOC is a process of continuous improvement to optimize flow through a system.
• Throughput, inventory, and operating expense are operational measurements used to determine if a change in the process will be productive or nonproductive.
• Throughput is the rate at which the system generates money through sales; inventory is all the money the system invests in purchasing things it intends to sell; and operating expense is all the money the system spends to turn inventory into throughput.
• The goal of TOC is to increase throughput, decrease inventory, and decrease operating expense.
• In a healthcare system, throughput is the rate at which a patient moves through a hospital or clinic; inventory is the number of patients at different stages in the hospital or clinic; operating expense is the cost of evaluating, diagnosing, treating, and discharging patients, plus all other costs associated with running the hospital or clinic.
• The goals in a healthcare system are faster dispositions, decreased time that a patient stays in the hospital or clinic, and more efficient treatment plans.
• A bottleneck resource is any resource whose capacity is equal to or less than the demand placed on it.
• TOC uses the 5FS to identify the bottleneck and improve flow through the bottleneck:
  Step 1: Identify the system’s bottleneck.
  Step 2: Decide how to exploit the system’s bottleneck.
  Step 3: Subordinate everything else to the decision made in Step 2.
  Step 4: Elevate the system’s bottleneck.
  Step 5: If the bottleneck has been broken, go back to Step 1.

NOTES

1. A note about these measurements: Throughput is technically the price paid for a product, less the variable costs that specifically depend on how much it costs to produce an additional unit—raw materials, for example. Fixed costs such as labor or rent fall under the category of operating expense. The throughput is expressed as a rate $/unit of time, where $ refers to margin made from sales and time refers to the period of time being examined. Additionally, as the practice of TOC matured, inventory became known as investment and included all assets that could be sold (e.g., equipment, buildings). Here again, we will focus on inventory as defined in the text.
2. We’ll leave out the part about the product being sold as a part of throughput (remember, it’s a rate at which the system generates money). Strictly speaking, throughput involves the money the healthcare system generates by treating patients, which is largely dependent on reimbursement by commercial payers or government payers. Practically speaking, flow in healthcare is not dependent on reimbursement, so in these discussions, throughput will simply be the rate at which a patient moves through the system.

REFERENCE