Improving Business Performance by Managing Constraints

C.R. Marshall, Ph.D.
Assistant Professor of Business Administration
University of Wisconsin - Stevens Point

Introduction

For at least two centuries, economists and management theorists have been introducing new ways for businesses to improve productivity. In 1776, Adam Smith recommended increasing the division of labor in production, using the manufacture of hatpins as an illustrative example. A century ago, Fredrick Taylor advised Bethlehem Steel to provide its laborers with shovels designed to handle the specific task at hand. The amount of iron ore, coal and ash moved by a worker increased from 16 tons per day to 54 tons per day. More recent managerial theories include Management by Objective, which seeks to motivate workers by involving them in the goal setting process; TQM, which involves everyone in the company in the quality of the finished product; and MRP and JIT which manage inventory to eliminate stock-outs and control inventory costs. Each new theory, when properly implemented, helps businesses improve their performance. The management theory discussed here is the Theory of Constraints, which improves business productivity by focusing the organization's attention on the part of the process that is most limiting to productivity increases.

Goldratt and the Theory of Constraints

Eliyahu M. Goldratt developed the Theory of Constraints (TOC). Goldratt, a Physics professor from Israel, first published the basics of TOC in 1984 in The Goal, a novel about a plant manager. TOC is a management philosophy that extends the concepts of Just-in-Time inventory control, and applies them to various aspects of management. This paper summarizes major aspects of TOC and discusses the application of these ideas in various work settings.

Application to the Service Sector

As the paper progresses, the discussion will constantly refer to inventory. A reader employed in the service sector may be initially inclined to set the paper aside, saying: "Inventory is such a small part of my business that this paper will be of little or no use to me." Before you do so, please take a moment to consider a broader view of inventory. In an academic setting, incoming freshmen are raw materials, enrolled students are work in process and graduating seniors are finished goods. In the insurance industry, contracts and claims that must be processed flow through the system much like materials flow through a manufacturing facility. In health care, the paperwork for admitting a patient follows a similar input-process-output path. Patients in for a series of tests can be thought of in the same manner. Efficiency gains from using better inventory handling methods lead to more effective capacity and happier patients. If you can expand your view of inventory to include paperwork and people or jobs processed in your industry, you will find that examining models of inventory management may help you to improve the performance of your organization.

The Importance of Continuous Improvement

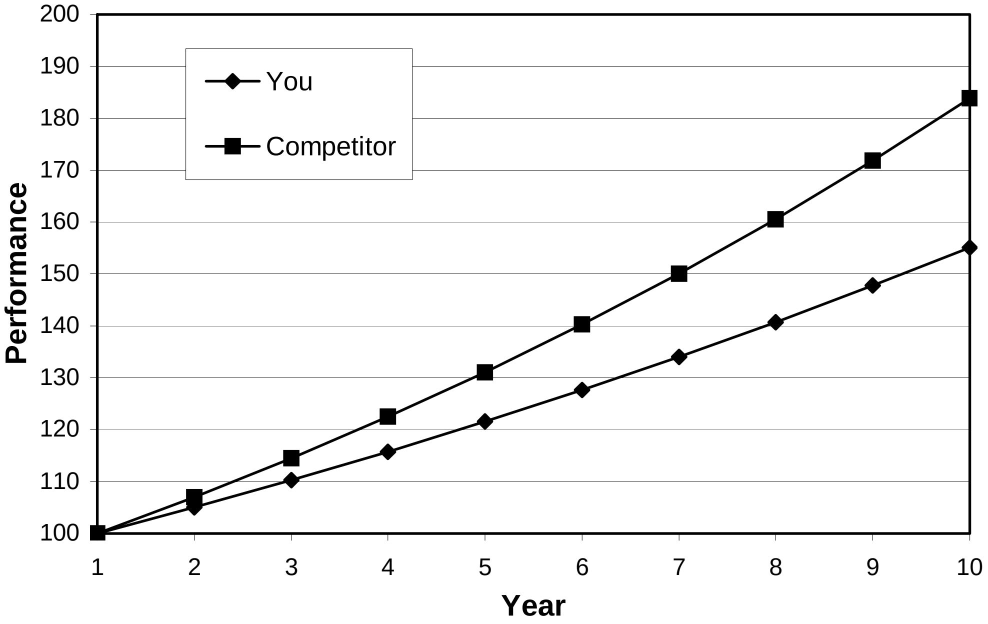
Competitive forces in our economy compel firms constantly to seek new ways to improve performance. Improvements in quality have not satisfied customers' demand for quality, but instead have led to even higher demands for quality. The same is true of product features, product diversity, on-time delivery, lead times and many other aspects of business performance.

In each case, it is the relative, not the absolute level of performance that dictates who gets the business. Let's look at an example where one of your competitors is improving at a slightly higher rate than you are. For this example, we will use an arbitrary scale where the current performance is measured as 100. Assume that both firms start off with equal current performance. Further assume that you are improving at 5% per year and that your competitor is improving at 7% per year. Table 1 and Figure 1 below show the effect of this difference over a ten-year period. In ten years time your competitor will be ahead by a substantial margin. Even though you started as equals, the difference is now 29 units on our arbitrary scale.

Table 1. Effect of a small performance difference over 10 years

Year	You	Competitor	Difference
1	100	100	0
2	105	107	2
3	110	114	4
4	116	123	7
5	122	131	10
6	128	140	13
7	134	150	16
8	141	161	20
9	148	172	24
10	155	184	29

Figure 1: Effect of a small performance difference over 10 years

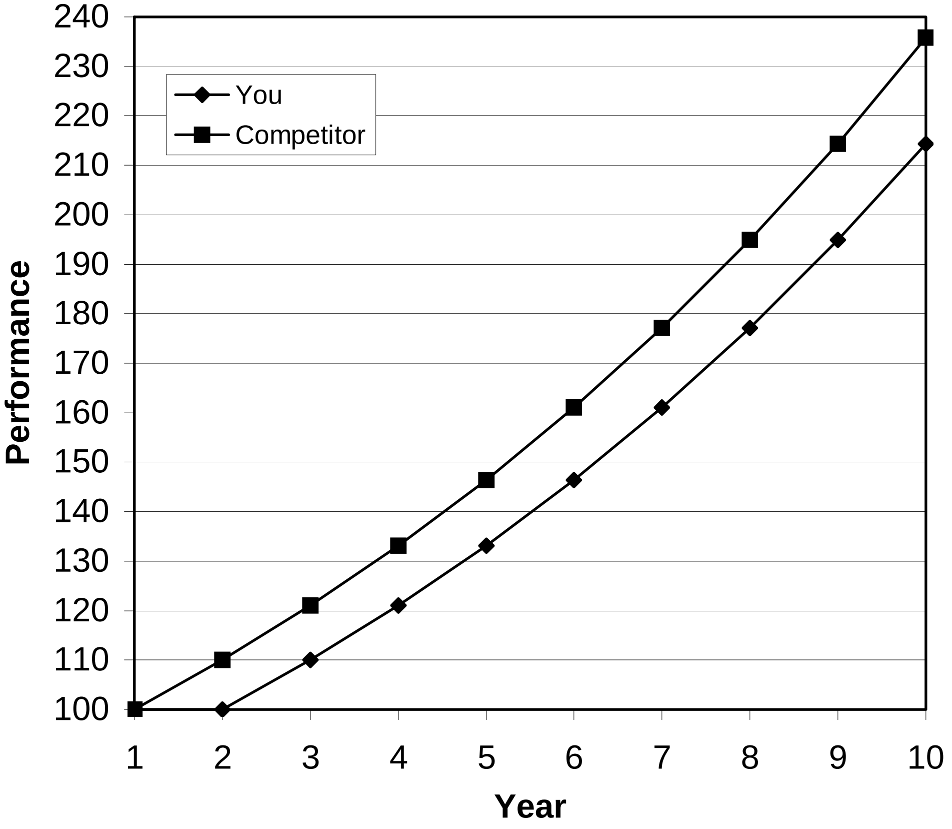


Now, let's assume that you are able to copy the industry leader and perform as well as they do. In order to benchmark your performance against a competitor, you must observe them and determine what they are doing and how they are performing. Table 2 and Figure 2 assume that the industry leader is improving their performance at 10% per year, and that you are able to match their performance one year after they achieve it. This is not quite as bad as the previous case. In 10 years you have gone from being equal to your top competitor to being behind them by 21 units of our performance scale. The gap will continue to spread.

Table 2. Effect of a performance lag over 10 years

Year	You	Competitor	Difference
1	100	100	0
2	100	110	10
3	110	121	11
4	121	133	12
5	133	146	13
6	146	161	15
7	161	177	16
8	177	195	18
9	195	214	19
10	214	236	21

Figure 2. Effect of a performance lag over 10 years



These two examples illustrate the importance of being the performance leader in your industry on those aspects of your product or service that your customers value the most. The implications for profitability are discussed below.

Goals and Constraints

Before determining the limits of a system, we must first determine the system's primary goal. If you are a for profit organization, then your primary goal is profit. The goal is to make money, now and in the future. All other goals are subsidiary to, and must be supportive of this primary goal. Operating efficiency is not an end but a means to attain higher profits. Customer satisfaction is vital to the ongoing viability of an organization. The reason we want satisfied customers is that we want them to return and give us more money. We want them to tell their friends about our organization so that those friends will give us money. Customer satisfaction is just a means to an end. Market share allows us to have more market power and possibly a stronger reputation, leading to higher profits.

Sub-Goals

Divisions, departments and individuals are often given specific goals on which they or their managers are evaluated. Sometimes the goals are specific. Sometimes they are vague. The list includes, but is not limited to, cost effective purchasing, production quotas, customer satisfaction, efficiency and hiring the best people. None of these sub-goals are the reason that your organization is in business. The company was not founded to purchase at the best price or to provide employment. These are the things that you do to help the company earn money.

Once you understand your objective, you can start looking at constraints. A constraint is anything that limits your ability to attain your objective. Every chain has a weakest link. It is the weakest link that constrains the department or organization.

Six Measures

Performance measures are necessary for an organization to determine if they are meeting their goals. We will start with three traditional measures of performance.

- **Net Income (NI)**: Profit remaining after expenses have been subtracted from sales.

NI is an absolute measure of performance. While it is important, you also need a relative measure.

- **Return on Investment (ROI)**: Income divided by the amount invested to produce that income.

ROI provides a relative measure of performance. A profit of $100,000 represents a good return on a $250,000 investment, but $100,000 of profit is inadequate if the investment necessary to produce it is $250,000,000.

- **Cash Flow (CF)**: The money available to pay current expenses.

Many firms with adequate net income and reasonable return on investment have gone out of business because of inadequate cash flow. If you have enough cash then cash flow is of lesser importantce. If you do not have enough cash, it is the most important of these three measures.

Within this framework Goldratt defines three additional measures of operational performance:

- **Throughput (T)**: The rate that the system generates cash through sales.

While production of finished goods inventory to fill a warehouse or the production of work in process to sit in front of the next step in the process keeps workers busy, it does not generate income and is not part of throughput.

- **Inventory (I)**: The money that the system has invested in purchasing things that it intends to sell.

In a traditional production setting this definition is clear-cut. In the service sector, where customers may be waiting to be served, this may take more analysis to become clear. Waiting time costs your customer money. If a series of medical tests takes a day to complete instead of a half day, your customer has paid an additional half day of his or her salary for the tests. While this expense does not show up directly in your accounting records, it increases the portion of your facility that must be devoted to waiting rooms and potentially decreases demand.

- **Operational Expense (OE)**: The money the system spends to turn Inventory into Throughput.

While the traditional measures are well understood by upper management, workers in the operations area often feel disconnected from the financial and accounting measurements. People at any level of the organization easily understand Goldratt's additional measures.

How do the new measures relate to the old ones? Will having workers focus on the new measures satisfy management's desires for performance in the areas of the old measures? Improving Throughput improves Net Income, ROI and Cash Flow. Improving Throughput, if it can be done without increasing inventory and without a disproportionate increase in OE, should make management happy. If Operating Expense can be reduced without reducing Throughput, then, Net Income, ROI and Cash Flow will improve and management will be happy. Reducing inventory, as long as it does not reduce Throughput, improves ROI. It also helps CF by having less money tied up in Inventory. It also reduces OE by reducing carrying costs, and therefore has an indirect effect on NI as well.

Herbie - A Constraint

To illustrate how jobs flow through a typical production process we will use a story adapted from The Goal. Imagine a troop of Boy Scouts on a 10-mile hike to a campground. If all of the scouts were of equal height with equal strides and all had the same level of fitness, the same load in their packs and the same constant walking speed of 2 miles per hour, we could start the hike with the scouts lined up in single file and arrive as a group at the campsite five hours later.

Unfortunately, the assumptions made here are far too numerous and rather unrealistic. Let’s add some realistic assumptions and see what happens to our five-hour hike. The first assumption will be that the scouts don't walk at the same speed. While the average speed of the scouts in the troop is 2 mph, some scouts average over 2 mph and others are slower. Lets name the slowest scout Herbie.

What will happen if the trail is wide enough for a faster scout to pass a slower scout? The faster scouts will move to the front of the line and the slower scouts will find themselves at the tail of the line. Once the scouts are in order from fastest to slowest, they will continue to spread out, with more and more space between each scout. How long will it take until the entire troop is at the campsite? The answer is that it will take longer than the 5 hours predicted in the previous paragraph. The time that it takes for the troop will be the time it takes for Herbie, our slowest scout to walk the ten miles. Herbie is our limiting factor, our constraint.

Now add two more assumptions. First, the path is narrow, so that a faster scout cannot pass a slower scout. Second, each individual scout sometimes walks faster than his individual average and sometimes slower. If the scouts are in order with the fastest scout at the front of the line to start, then the hike time will be the same as above. The line will spread out. The first scout will arrive in less than five hours. The troop will not be complete until Herbie finally arrives sometime past the five-hour mark.

What if the scouts start out in random order instead of being perfectly ordered by their hiking speed? Faster scouts in the middle of the troop will find their performance constrained by the slower scouts in front of them. This means that the average speed of all of the scouts except the one at the front of the line will fall.

Now consider what happens if one of the scouts in front of Herbie stops to adjust his pack or tie his shoe. When Herbie is walking above his average speed he can keep up with the scout in front of him. When the scout in front of Herbie stops, Herbie has to stop and wait, and his average speed falls. Anything that slows Herbie down lengthens the time it takes to complete the hike.

Note that in the previous examples where the scouts were ordered from fastest to slowest, the space between the scouts increased as the hike progressed. Once they are sufficiently spread out, a scout stopping to adjust his pack will not have any effect on the speed of the scout behind him. Early in the process they are still bunched together, and any scout that slows down may cause Herbie to slow down, and may therefore increase the total time for the hike.

In our analogy, the first scout is the first step in the production process. Each subsequent scout represents the next step in the process. The trail in front of the first scout represents raw materials. The space between scouts is work in process. The last scout is the last step in the process. The trail behind the last scout is finished goods.

Statistical Fluctuation and Dependence

Compare the hike to most production and service processes. If there is variation within and between the steps in the processes in your organization, then you have statistical fluctuation. This problem is well discussed in the quality literature and reduction in this fluctuation is a necessary condition for just-in-time inventory systems.

If the steps in the process must be done in a particular order then the process steps have dependence. Wood must be sanded before it is painted, the value of an insurance claim must be determined before the claim check is processed, and patient insurance status should be confirmed before a hospital room is assigned. The flow of work into any station depends on the timely completion of the work in previous stations. This is the narrow trail / no passing assumption in our analogy. In order for any scout to move forward, the scout in front of him has to have already moved forward. Steps that are later in the process must wait for all earlier steps to be completed.

Five Steps

Here are the steps in the improvement process

1. Identify the system's constraints.

In this step, the manager must determine which of the system's constraints is the primary limiting factor. The manager must identify Herbie. Managers generally know where the bottlenecks are in their system. Attention should be focused on the particular bottleneck that is the tightest constraint.

2. Decide how to exploit the system's constraints.

What can be done to make the bottleneck process more efficient? Do the employees need more or better training? Can the task be redesigned or automated? Can workers be rescheduled to reduce or eliminate downtime and idle time? Can workers be reassigned to this task during peak load periods?

3. Subordinate everything else to the above decision.

Every part of the process that is not a bottleneck by definition has slack. These areas should manage their workflow, using the available slack if needed to support production in the constrained department. The constrained department should never be idle because of the actions of some other part of the organization. An hour lost at the constraint is an hour lost for the system.

4. Elevate the system's constraints

If sufficient work is done on the constraint, whether it is a redesign of the process, additional training, additional personnel or equipment, combined with the support of upstream departments, the constraint can be lifted to the point that it is no longer binding.

5. If a constraint has been broken, go back to step 1.

When performance in the primary problem area has been sufficiently improved, it will no longer be the problem area. Something else will now be the binding constraint.

When searching for potentially binding constraints, there are two things to look for. First, look for large piles of inventory waiting to be processed. Remember that this inventory may be in the form of a constantly overflowing inbox or a waiting room that tends to be standing room only. Second, consider departments that are constantly demanding more resources, such as additional space, equipment or manpower.

Another thing to remember about constraints is that they are rarely the result of insufficient space, equipment or manpower. They are generally the result of constraining and often outdated policies. An example can be found in an engineering department that was overloaded and unable to produce design changes in a timely fashion. The complaint was that the computers that ran their CAD system were slow and out of date. The engineers were in fact busy all of the time, and each had 2 or 3 jobs on their desk that they were working on simultaneously. The problem was not the speed of the computer system but the policy that allowed jobs 2 and 3 to be started before job 1 was finished. Completion of job 1 was postponed while job 2 was being started. Completion of job 2 was postponed because of work on job 3, and so on. By not issuing a new job to engineers until the previous task was completed, the change in policy effectively reduced turnaround time by roughly one-half.

Some Possible Solutions

Herbie Leads

If you want the scouts to arrive together and to arrive as quickly as possible, one possible alternative is to have Herbie lead the troop. When Herbie is slower than his average he slows down the arrival by exactly as much as in the previous example. When he moves faster than his average speed, the whole troop can move faster because all of the scouts behind him can catch up, at least eventually. With no one in front of Herbie to constrain his performance, he is able to move at his average speed.

Production Schedule (Drums)

There are steps in most processes that depend on other processes having already been completed. If there is dependence then Herbie cannot be moved to the front of the line. The operations that follow the constraint should have enough slack capacity to keep up with the pace of the constrained resource. The problem lies with the operations that precede the constraint. The scouts that are ahead of Herbie are not constrained by his slow pace. They are free to proceed at their own pace. Like the fast scouts at the front of the line who move ahead and produce gaps, these operations will move ahead and produce excess WIP inventory. The carrying cost on this inventory increases OE, reduces NI, ROI and CF.

A possibility for keeping the front of the line from running away from the rest of the operations is, in effect, a drum beating a cadence that the constraint can keep up with. The drumbeat in a manufacturing setting is the production schedule, which dictates when and what material is supposed to be processed by what resource. Once you realize that the troop cannot move faster than the constraint, it becomes obvious that the production schedule must be dictated by the abilities of the constrained resource. If every production resource can produce to match the production schedule, then this system should work. If they cannot, then there will either be delays that reduce throughput or increases in work in process inventory.

Sometimes the drum system is used in a push inventory system, where raw materials are released into the system to keep workers busy and keep efficiencies high on each resource in the system. In this case the cadence is not set to the slowest worker, but is set at or above the average. Expeditors and additional managerial attention are often needed to push work through the slower workstations. You can think of this as a Just-in-Case system. Work in Process inventory is high so that down time on any portion of the system does not endanger current production. High levels of inventory reduce the company's ability to respond to changing customer demands. While reported efficiencies are high, this system has a detrimental effect on NI, ROI, CF and OE, and therefore threatens future Throughput.

Assembly Lines, Balanced Lines and JIT (Ropes)

The next possibility is a rope connecting each scout. This is effectively what you get with an assembly line or a production line where attempts have been made to balance the capacity of each workstation. The speed of the line beats the cadence and the structure of the line connects the workers to each other. The same system also describes just in time inventory systems. Here the cadence is set by market demand for finished goods. Transfer batch sizes are low, as is WIP inventory. Since production is driven by demand rather than warehouse capacity, the material produced is Throughput and not stored finished goods.

The problem with this type of system is the existence of statistical fluctuations. With minimal work in process, any problem that occurs at any point in the process can bring the entire system to a halt. Successful JIT systems often require years of work to reduce the variability that naturally occurs in the process. While this may have a focusing effect for managers interested in solving production problems, it has a potentially devastating effect on current throughput.

Drum-Buffer-Rope (DBR)

What we need is a system that has low inventory and avoids downtime. The Theory of Constraints literature suggests a system called Drum-Buffer-Rope

(DBR). In this system Herbie sets the cadence in that the production schedule is determined with the goal of matching the capacity of the constrained resource. The rope connects the constrained resource to the first resource in the process. The production schedule releases material to the first operation at exactly the rate that the constrained resource can process it. Therefore neither the first resource nor any other resource that precedes the constraint can produce excess inventory.

A problem with any of these workstations in a Just-in-Time system causes Herbie to shut down and reduces throughput for the system. What is needed is a buffer of work in process inventory that will allow the earlier processes to catch up before the constraint runs out of work. Lengthening the rope connecting the constraint to the first process creates the buffer. These processes are constrained to run at the same speed as Herbie but are allowed to run slightly ahead.

Downtime on processes that follow the constraint is not a problem because they have sufficient excess capacity to catch up with the constrained resource. There should be no build up of excess WIP inventory for these processes, again because they have more than enough capacity to process everything that Herbie hands them.

A Simple Example

Consider a process that has three sequential steps performed by three departments. Step A requires 9 minutes to complete. Step B requires 10 minutes. Step C requires 8 minutes. The capacities of the departments are 6.67, 6, and 7.5 units per hour respectively. What is the capacity of the organization? It is the capacity of the slowest department. Improving the productivity of either department A or C has no effect on organizational capacity. Within limits, down time and idle time in departments A and C have no impact on capacity. On the other hand, every minute of downtime or idle time in department B reduces throughput by 1/10th of a unit. Idle time in B reduces productive capacity for the entire system.

First suppose that the system in place is the simple drum and the cadence is set at the average capacity of the three machines, around 6.7. Department A will be working at capacity. Inventory will pile up in department B's receiving area at a rate of 0.67 units per hour. Department C will process everything that they receive, but will be in trouble with management for low efficiencies (6/7.5 = 80% of capacity).

Suppose we correct the cadence and set it at the capacity of B. This process is now running as a balanced assembly line with inventory arriving just in time. Every 10 minutes A receives enough raw materials to make one unit. A inspects the unit in the last few seconds of the process, and hands the unit to B who processes it and hands it to C. What happens if A discovers that they have produced a defective unit? B and C have to wait while A produces a replacement. We have lost 9 minutes of capacity for the entire plant. We will have a similar problem if a supplier delivers the material late, if A's equipment has a mechanical failure, if A is late getting back from lunch, or if natural variation causes A to produce at a rate slower than average for any reason.

Based on historical information, we can estimate the magnitude of problems that department A might encounter, and we can calculate how much work in process inventory department B needs to have in their receiving area to sustain them while A catches up. For this example assume that 3 units are sufficient for A to catch up from any problems that they might encounter. We would allow department A to produce at their maximum capacity until they had accumulated 3 units of work in process. At that point they would have filled the buffer and hit the end of their rope. They would now receive only enough raw materials to produce one unit every 10 minutes. Whenever the buffer falls below the desired level, material is released to allow department A to replenish it to a safe level of work in process.

Conclusion

Both a pure drum system and a pure rope system have the same goal: the efficient operation of the system. They differ in their underlying assumptions. The pure drum system wants WIP inventory to protect against down time. Inventory is pushed into the system to try to keep all resources engaged. Each step in the process tries to operate at its locally optimal level. The pure rope system pulls inventory through, and seeks to reduce WIP inventory to eliminate potential quality and dependability problems hidden in large inventory pools, and to eliminate costs associated with carrying inventory and with slow response times to changing customer demands. The Theory of Constraints approach accomplishes the goals of each of these systems using the Drum-Buffer-Rope method. Inventory is accumulated where it is needed to avoid system down time, and eliminated where it is not needed to lower cost and improve response to customer demand.

Readings

Goldratt, Eliyahu M. and Jeff Cox, (1984) The Goal, North River Press, Great Barrington, MA

Goldratt, Eliyahu M. and Robert E. Fox, (1986) The Race, North River Press, Great Barrington, MA

Goldratt, Eliyahu M. (1990) What is this thing called Theory of Constraints?, North River Press, Great Barrington, MA

Goldratt, Eliyahu M. (1994) It's Not Luck, North River Press, Great Barrington, MA

Goldratt, Eliyahu M. (1997) Critical Chain, North River Press, Great Barrington, MA

Goldratt, Eliyahu M., Eli Schragenheim and Carol A. Ptak, (2000) Necessary But Not Sufficient, North River Press, Great Barrington, MA

Scheinkopf, Lisa J., (1999) Thinking for a Change: Putting the TOC Thinking Process to Use, St. Lucia Press, Boca Raton, FL

Schragenheim, Eli, (1999) Management Dilemmas: The Theory of Constraints Approach to Problem Identification and Solutions, St. Lucia Press, Boca Raton, FL