BLIND FAITH: THE HIDDEN COSTS OF CAPACITY OVERUTILIZATION

This article examines whether there are limits to the utilization of executional drivers, specifically capacity utilization.

CJ MCNAIR-CONNOLLY AND CHARLES R. THOMAS, JR.

In 1993, John Shank and Vijay Govindarajan launched a new discipline in the strategic accounting literature: strategic cost management.[1] While receiving good press at the time, it has proven more difficult than expected to actually test the model in realistic settings. Blending three different streams of strategy research — value chain analysis, strategic positioning analysis, and cost driver analysis — the theory set forth by these two authors moves cost out of the zone of operations and into the strategic domain. The core idea of the theory is that there is a range of structural and executional cost drivers that management can manipulate when faced with a strategic challenge. These cost drivers operate differently when an organization chooses a cost versus differentiation strategy.

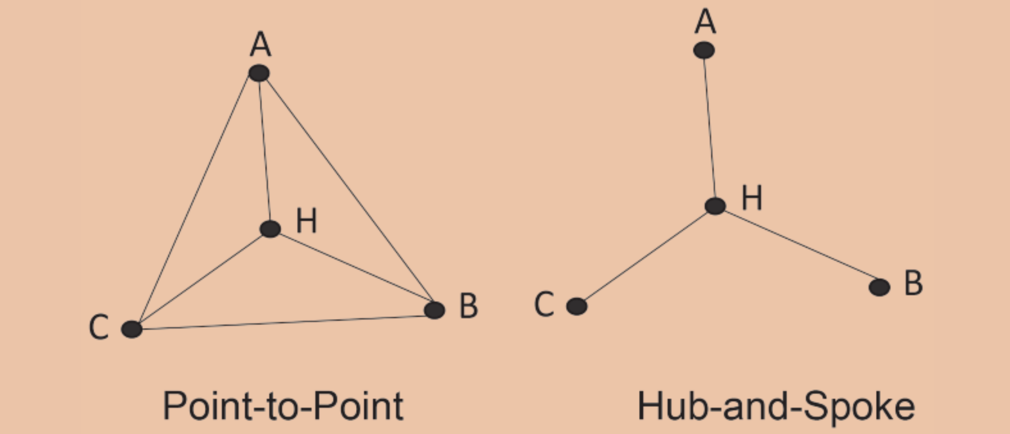
There are five main structural cost drivers noted in the model: scale (size of investment); scope (degree of vertical integration); experience; technology; and complexity (breadth of product line offerings). Each of the choices made in terms of structural cost drivers impacts the final product cost. Structural cost drivers, then, represent the constraints under which the business has chosen to operate. The amount of capacity, or the organization's scale, is a dominant aspect of the structural cost drivers that management has to manage.

Executional cost drivers, on the other hand, deal with the ability of the organization to execute its strategies within its structural constraints. For executional cost drivers, the two authors argue that more of the driver is always better. The executional cost drivers noted include workforce involvement, total quality management, capacity utilization, plant layout efficiency, product configuration, and the exploitation of linkages with the organization's customers and suppliers. Capacity utilization differs from structural capacity because it reflects decisions on how management uses the capacity it has purchased. It can be argued, however, that an organization that operates too

CJ MCNAIR-CONNOLLY, Ph.D., *is an internationally recognized expert in cost management. She has authored 10 trade books and numerous articles on various aspects of the relationship and development of cost management and the new technologies that define modern management practice. Holding an MBA and Ph.D. from Columbia University, Dr. McNair-Connolly is a retired professor of accounting from the U.S. Coast Guard Academy.*

CHARLES R. THOMAS, JR., Ph.D., *is associate professor of accounting at Tarleton State University in Texas. A CPA, he holds a Ph.D. from The University of Texas at Arlington and is an active CMA. Dr. Thomas has served as director of financial planning and analysis at Southwest Airlines and director of Ecole Hoteliere de Lausanne's executive MBA program.*

EXHIBIT 1 Point-to-Point Network Versus a Hub-and-Spoke Network



close to the limits of its potential capacity utilization is more exposed to the negative impact of operational and strategic disruptions. This potential offsets the notion that utilizing more of the available capacity of an organization is always a recipe for superior performance.

What remains as a question, then, is whether there are limits to the utilization of such executional drivers as capacity utilization. Specifically, if an organization operates too close to the physical limits of its structural capacity, does it not face an exponentially growing list of potential problems that could become a smoldering or acute crisis? In other words, does a snowball effect, or a growing list of problems and crisis events, begin to take place as an organization moves toward the outer limits of its available capacity?

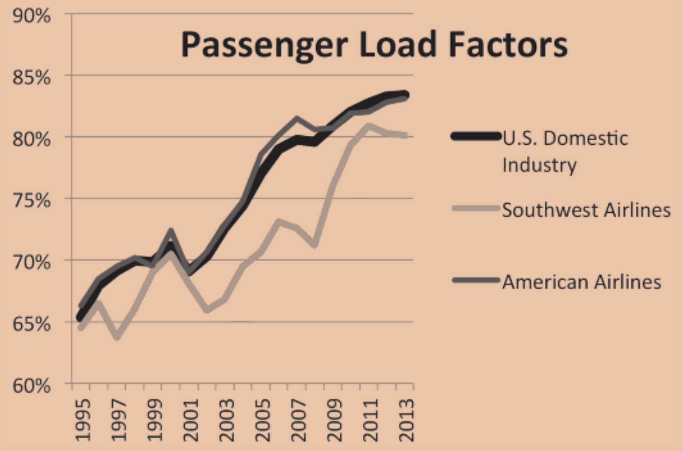
In this article we will explore the role played by escalating marginal costs of disruption as capacity utilization moves beyond specified limits. Seeking to identify both the more easily measured and less easily measured costs of capacity overutilization, the role of capacity utilization in an airline is used to explore some of the limiting features. The goal of this article is to overturn the notion suggested by Shank and Govindarajan that more is always better when it comes to executional cost drivers such as capacity utilization. In fact, overutilization can rob an organization of its flexibility to respond to normal problems of daily business, turning them into crises that can negatively affect the organization's ability to meet and exceed customer expectations. Let's start by laying the groundwork for the concept of capacity in the airline industry.

Airline capacity

Capacity in an airline setting is composed of a complex blend of assets and people. An airline creates a dynamic system of people, facilities, aircraft, and other equipment. Aircraft are scheduled to transit between airport stations that serve as nodes in a network that can be thought of as arcs connecting the nodes. First, an airline has to choose a network structure. The two most common choices that are made are between a hub-and-spoke (HS) design and a point-to-point (PP) flying network. Exhibit 1 shows the difference between these two approaches, using a simple four-node network for illustration purposes. What you can see is that in a PP network the emphasis is on directly connecting the physical nodes (airports) in the system, while in the HS network a traveler has to change planes at the hub location in order to make the connections between airports A, B, and C.

When operating an HS design, airline companies seek to concentrate their flights both spatially (through the hub) and temporally (flying waves of flights

EXHIBIT 2 Passenger Load Factors for Southwest Airlines Versus American Airlines and the Industry



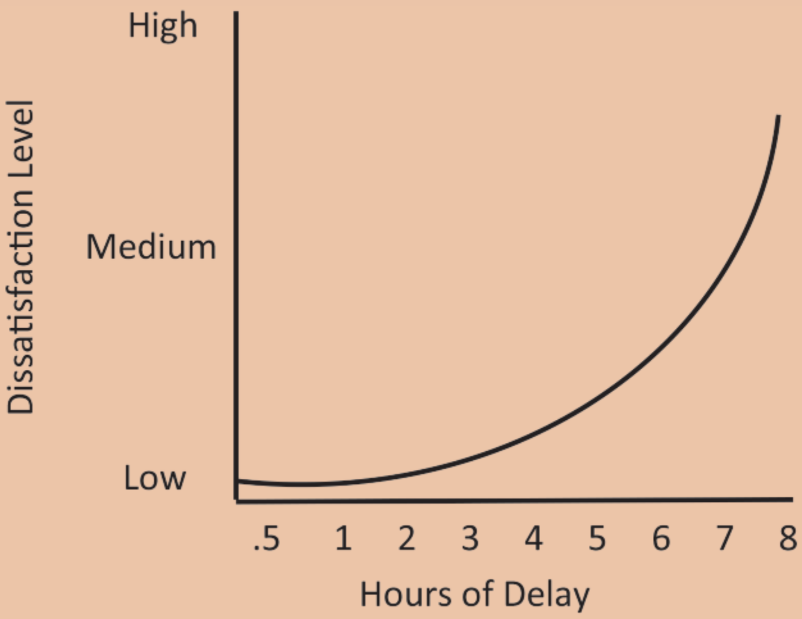
that emphasize connecting passenger routes). While there can be some level of spatial concentration in a PP network, there is no attempt to link the flights temporally. Each aircraft is scheduled to carry out its route with little coordination with other routes. If passengers are connecting flights within the network, it is simply coincidental to the true focus of the network operations; the goal is to maximize PP direct itineraries.

In choosing an HS design, the airline is emphasizing economies of scale: It purchases aircraft with different seating capacities (large or small) based on the projected traffic between one of the spoke airports and the hub airport. The logic behind scale economies in the airline industry is that the airline attempts to match the size (seat capacity) of the aircraft it uses to the projected traffic on a specific arc, or connecting link, between two airports. What results in an HS network is a number of small aircraft being used to connect network nodes (airports) to the central hub. Large aircraft are used in an HS design only where the projected traffic justifies its use. With this set of choices for the HS design, then, comes complexity in the form of multiple types of aircraft that have to be scheduled and maintained. In the PP network, there is a tendency to use one medium-sized aircraft, such as the Boeing 737 line, for all of the flights. Load factors are manipulated in PP networks by altering the frequency of flights between nodes. This simplified structure allows for cost savings, something the PP network offers as an alternative to the economies of scale pursued in HS designs.

Unfortunately, due to the tendency to link their flights temporally, HS designs end up with a significant level of underutilized resources. At hub airports, peak demand to handle waves or "banks" of flights exchanging travelers dictates capacity levels required of people, facilities space, and equipment. At spoke airports, flying times required for arrival at hubs during flight waves determine when demand peaks occur. HS designs thus face a significant cost in terms of standby capacity, which is one of the most expensive forms of capacity waste. In order to accommodate the exchange of travelers, aircraft are scheduled to wait rather than continue on their routes. Uncertainty introduced by occasional schedule disruptions leads airline managers to buffer flight schedules, which, in turn, leads to more aircraft and crew waiting. With the ability to match the scale of aircraft used to the routes and because flight waves maximize connecting-itinerary opportunities, there tends to be high utilization of the seat capacity on the aircraft itself. In looking at Southwest Airlines, a PP airline, we see historically lower seat utilizations than its HS competitors (see Exhibit 2). American Airlines is one example of an airline that uses an HS network design. As can be seen from the exhibit, its seat capacity utilization is higher than that of Southwest Airlines. In fact, most of the full-service major carriers employ either a single- or multi-hub design.

In summary, there are three primary components to the capacity of an airline: the number of nodes (airports) served by the network, aircraft time, and the number of seats available on an individual plane that is flying within the network. All three aspects of airline capacity have the potential for capacity waste: underutilized nodes with significant standby capacity due to flight schedules, aircraft waiting for travelers, and unoccupied seats on aircraft flights.

EXHIBIT 3 Customer Dissatisfaction with Flight Delays

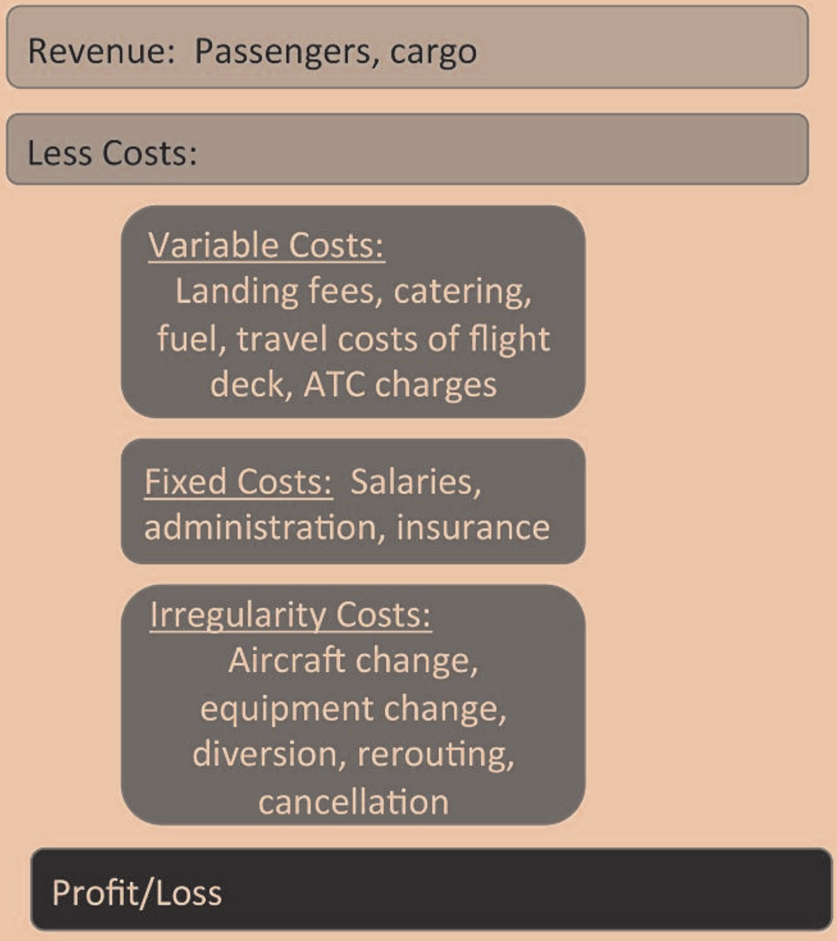


Passengers on an airline make the first choice regarding the airline's network when they decide to fly with one carrier versus another. Everything else being equal, passengers prefer nonstop or direct (no aircraft change) itineraries. If no such itineraries are offered between the cities being linked by the passenger, the PP network loses its competitive advantage for the traveler and becomes one of many airlines that the passenger can choose to fly. At this point, the price of the airline ticket for the desired route becomes a driving factor in airline choice. Some airlines have used other enhancements, such as generous frequent flyer programs, first-class seating availability, and airport clubs, to gain a greater share of this connecting passenger traffic.

What is of specific interest is the question of how "full" an aircraft should be on a specific route. Should the goal be to fill every seat, resulting in the overselling of capacity on the aircraft at certain times? Clearly the low marginal cost of filling an additional seat makes such a move look the most promising for the airline, but it opens the airline to the impact of disruptions. If bad weather, for instance, causes the cancellation of one or more flights, there simply is not enough available seat capacity in the system to clear the passengers through to their final destination. It can be argued, then, that overutilization of a plane's capacity on a regular basis leads to a situation in which normal operating problems, such as losing the use of a plane due to maintenance problems or dealing with weather problems in a city or region, becomes a crisis that can wreak havoc with the financial, reputational, and relational subsystems of the airline. Customer dissatisfaction soars under conditions of long, unplanned travel delays, causing passengers to switch their buying behavior to another airline in response to the lengthy delays.

What are the sources of disruption delays for an airline? Aircraft maintenance, crew problems, and other circumstances within airlines' control lead to about one-fourth of all delays. Airport operations, air traffic control, heavy traffic volume, and weather conditions cause another one-fourth of delays. Late aircraft cause about one-third of delays as travelers await aircraft and crews from flights previously delayed. In the HS design, there are delays that occur because planes are being held for connecting pas-

EXHIBIT 4 Profit and Loss Calculation for a Single Flight



Revenue: Passengers, cargo

Less Costs:

Variable Costs: Landing fees, catering, fuel, travel costs of flight deck, ATC charges

Fixed Costs: Salaries, administration, insurance

Irregularity Costs: Aircraft change, equipment change, diversion, rerouting, cancellation

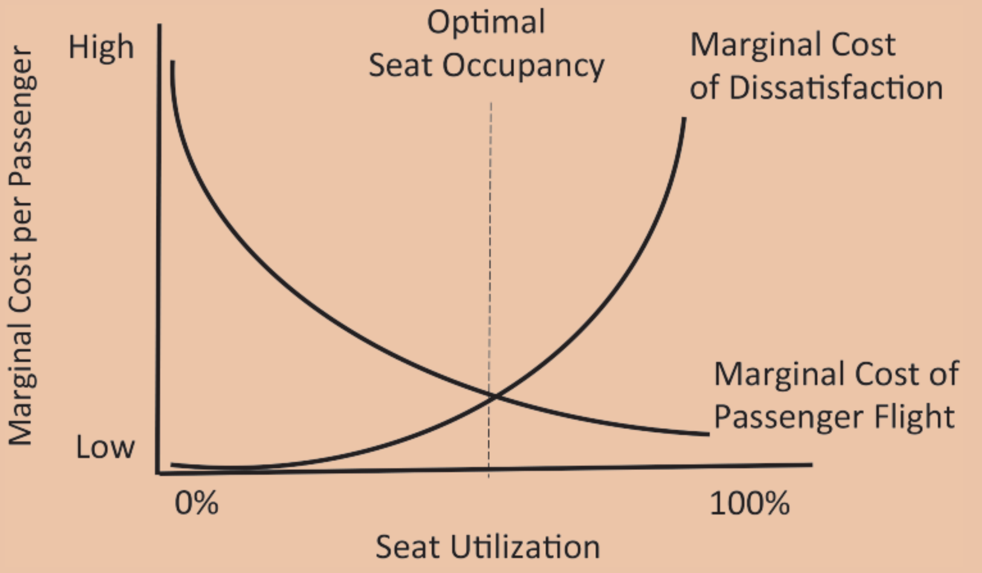
Profit/Loss

sengers and the associated baggage and freight. There are also instances in which mandatory security holds take place, when all baggage and passengers are required to deplane for inspection. Even with good scheduling software, the airlines often face delays as flight and cabin crew rotation results in a gap in the capability to fly a specific flight. Finally, there can be strikes and other actions that delay a plane from its specified route.[2]

The problem that underlies this extensive list of potential delays is the fact that once a scheduling disruption of any type occurs, a snowball effect takes place. The delays of one aircraft roll through the network, impacting the performance of other flights that share either the physical airport assets, the aircraft, or the flight deck or cabin crew. It has been found, in fact, that the longer the delay of one aircraft, the larger the impact on the performance of the whole network.[3] This impact does not grow linearly; it accelerates following an exponential path of increasing disruption in the system. The resulting delays add to traveler frustration, especially if a specific flight is cancelled due to unplanned delays that cannot be remediated.

The level of customer dissatisfaction with delays on a specific trip has been modeled in the transportation literature as suggested by Exhibit 3. As can be seen, customer dissatisfaction grows exponentially as the length of the delay grows. With an HS design, this dissatisfaction is magnified by the fact that flight delays from the outer nodes of the network can result in missed connections at the hub airport,

EXHIBIT 5 Marginal Costs of Seat Utilization



which, in turn, increase delay duration. While airlines cannot prevent the daily problems caused by aircraft maintenance and weather delays, if they choose to run all flights at a fully occupied seating level, the delays for individual customers are increased: They have to wait for an unoccupied seat to become available on the entire route before they can once again continue their trip. Since this added delay can lead to significant levels of frustration, the passengers can become biased against the airline. The result can be a financial, reputational, and relational crisis for the airline: Its passengers go elsewhere with their business.

Marginal costs of occupancy choice

In thinking about where the optimal point is in terms of seat occupancy rates on individual flights, two forms of cost have to be considered. The first cost, the marginal cost of flying another passenger on a planned flight, decreases as more passengers are added. Once an airport station has been established, the aircraft has been acquired, the personnel has been hired, and a schedule of flights has been developed, most of an airline's costs have been committed, including the following:

- costs incurred at the departure airport;
- costs incurred at the arrival airport;
- pilot and flight attendant costs; and
- fuel and oil consumed.

None of these costs will vary much in relation to aircraft seat occupancy. For example, consider fuel. For a specified aircraft, fuel consumed varies with flight distance and payload. But changes in payload, for many modern commercial aircraft, lead to small changes in fuel consumed. For a 1,000-mile flight in a Boeing 737-700, an additional 1,000 pounds of payload (four to five passengers and their bags) will increase fuel consumption by only about 10 gallons, about $20 at today's prices.

Because these costs do not increase much as passenger counts rise, increasing passenger counts is quite attractive; the contribution margin increases and average cost per passenger decreases. Exhibit 4 details the profit and loss calculation for a single flight.[4]

Offsetting the marginal cost of a flight being considered in this article is the marginal cost of dissatisfaction of the customer as seat occupancy increases. Specifically, these marginal and often hidden costs include:

- decisions by disrupted passengers to switch airline carriers in the future (loss of revenue);
- reputational loss as customers complain to the press or to other potential passengers;
- stress between ground agents trying to find flights for disrupted passengers;
- stress felt by passengers themselves, leading to job dissatisfaction and relational losses;
- costs of hotel rooms, meals, and toiletries incurred when an airline accommodates disrupted passengers during the unplanned wait time;
- costs arising from handling late bags, including added labor and shipping charges;
- gate and related airport charges as flights get delayed and stationed either at the gate for extended periods or on the tarmac, coming back to the gate when connecting passengers and cargo arrive; and
- aircraft flight and cabin crew disruptions that can result in job dissatisfaction and excessive stress for employees.

These hidden operational, relational, and reputational costs clearly increase as the length of the flight delay increases, which was suggested by the dissatisfaction curve. We can now bring these two types of cost together, yielding the results captured in Exhibit 5.

What the exhibit suggests is that optimal seat utilization falls somewhere below full occupancy, probably around the 75 percent level of usage. While this leaves the airline with less incremental revenue, it also provides it with the capacity required to accommodate passengers whose flights have been disrupted. In other words, the decision to purposely leave some capacity available to deal with the impact of daily disruptions prevents these daily problems from growing into full-blown crises. This suggests that there is a downside to deciding to fill all available seats on every plane: The airline loses the necessary slack that allows it to effectively respond to the impact of daily operational problems. When the hidden costs of overutilization of capacity are considered, it appears that there is a point at which diseconomies arise from further scheduled utilization of airplane seats.

OVERTURNING SHANK AND GOVINDARAJAN'S ARGUMENT THAT MORE IS ALWAYS BETTER WITH EXECUTIONAL COST DRIVERS, IT APPEARS THAT WHEN THE HIDDEN COSTS CAUSED BY A DISRUPTION IN THE OPERATIONS OF A SYSTEM ARE FACTORED IN, THERE ARE LOGICAL LIMITS TO HOW MUCH SEAT CAPACITY SHOULD BE PLANNED FOR UTILIZATION.

Looking to other industries

The situation facing airlines has been used to illustrate the argument that there are levels of capacity utilization that actually result in negative outcomes overall. Overturning Shank and Govindarajan's argument that more is always better with executional cost drivers, it appears that when the hidden costs caused by a disruption in the operations of a system are factored in, there are logical limits to how much seat capacity should be planned for utilization.

In the engineering literature, the argument is made that a system should be designed with the intention to only utilize 80 percent of the available capacity to allow for the flexibility to deal with disruptions. While prior arguments have been made that this creates a form of rate-based waste of capacity, when the hidden costs of disruptions are considered the decision to plan to utilize less than the maximum amount of capacity makes sound economic sense.

An example of where a decision to utilize less than 100 percent of the available capacity makes sense is in retail. If the checkout counters are designed so that all of them have to be running in order to keep up with demand from customers, it does not take much imagination to see that if disruptions, such as difficult transactions or delays for price checks or related activities, take place, the queue of customers begins to grow. Since all of the registers are utilized, there is no way for the retail store to deal with the queue; they simply have to wait for the lines to clear as service demand slips below service capacity. For customers, there is increased frustration as wait time escalates. This can lead to a decision to shop in other stores where they can be more rapidly served.

Leaving retail, one can enter the world of manufacturing. Here the organization faces potential disruptions from machinery breaking down or supply shortages, bringing production to a halt. If the factory is operating under a condition in which every available hour is needed to meet customer demand, delays begin to take place in delivery times quite rapidly. This once again leads to customer dissatisfaction that can turn a machine breakdown (a problem) into a financial, reputational, and relational crisis with the plant's customers. The hidden costs of crossing the problem–crisis threshold underscore the argument that some capacity needs to be set aside to allow the organization to effectively deal with common disruptions to operations.

The list of industry examples could go on. The point being made is simple: The costs of a small level of underutilized capacity is less than the marginal costs caused when disruptions impact an organization's customers. Only in a problem-free (disruption-free, with smooth, stable demand) world would total capacity utilization make some sense. Daily disruptions can only be prevented to a certain extent. If machinery is taken offline for maintenance before a problem occurs, the organization is recognizing that the cost of the lost production is less than the cost of disrupted service should the machine go down when production is planned.

Summary

It has long been argued in the capacity literature that more utilization of available capacity is better because it drives the marginal cost of another unit of output down as fixed costs are spread over more units. This argument becomes a problem, though, when encountering the fact that disruptions occur in every type of business. As we saw with the airline, both maintenance and weather problems are a daily challenge for keeping flights on schedule. If all of the seats on all of the flights are fully filled, the airline has no flexibility for dealing with the impact of these daily disruptions. The result is dissatisfied customers, who bring with them a range of hidden costs that can balloon into an acute crisis if not properly dealt with early in the process. It is important, then, for organizations to try to put a value on the marginal costs that are caused by disruptions when considering how much of their capacity should be planned for optimal utilization. This optimal point will fall below 100 percent, regardless of the fixed cost nature of the organization. Pushing the limits of capacity opens the organization to hidden costs as customers are impacted.

In the future, studies could be done that attempt to put actual monetary values on the hidden costs of capacity overutilization. It would be interesting to study industries, such as papermaking, for which 24/7 operations are the norm. How do these companies deal with daily disruptions and how do they quantify the impact they make on overall output? Do they plan for total utilization but factor disruptions into the delivery schedules so customers are not affected? This is just one strategy that they might use to keep daily disruptions from turning into customer-based crises. In the end, one thing is clear: Managers have to factor in the hidden costs of disruptions when choosing how much of their capacity to place in scheduled, planned utilization. A failure to capture the impact of disruptions can lead to escalating costs and negative relations with customers. ■

NOTES

[1] Shank, J. and Govindarajan, V., *Strategic Cost Management: The New Tool for Competitive Advantage.* (New York: The Free Press, 1993).

[2] "The costs of delays and cancellations: Analysis and means for cost reduction," M2P Consulting, presentation at AGIFORS, Dubai 2006.

[3] Zou, B. and Hansen, M., Impact of operational performance on air carrier cost structure: Evidence from US airlines, *Transportation Research Part E: Logistics and Transportation Review* 48, no. 5 (Sept 2012): 1032–1048.

[4] *Op. cit.* note 2.