Reprinted from the California Management Review, Vol. 43, No. 4.
by permission of The Regents.

Tradeoffs in Responses to Work Pressure in the Service Industry

Rogelio Oliva

Increasing class size throughout the public education system and teachers' burnout; rising airline load factors and airport utilization rates reaching historical highs; long waiting times for emergency care and shorter hospital stays after surgery—these are all symptoms of systematic imbalances in demand and supply in service-providing organizations. The service sector represents over 70 percent of the U.S. economy and provides almost three out of every four jobs in America. Studies consistently report that the main determinant of customer satisfaction, even when purchasing sophisticated products, is the level of service obtained from the supplier. Furthermore, managers of service firms are also service consumers, and they surely appreciate the frustration of poor or unfair treatment in a service transaction. It seems paradoxical that services are recognized as critically important to our economy and yet episodes of poor service proliferate. The complaint of poor services has regularly been picked up by the popular press over the last twenty years.[1] Even during a booming economy, we still do not seem to get services right: the American Customer Satisfaction Index for services fell in 1999 to 69.4, down five percentage points from its 1994 value.[2]

While imbalances in supply and demand eventually translate into consequences for customers (long waiting times or reduced attention from service personnel) and for the profitability of service enterprises (excess capacity or reduced revenues due to unsatisfied customers), they first manifest as work pressure for the servers. Service personnel perceive work pressure as the difference between the amount of work that can feasibly be done and the amount of

Support for this research has been provided by the Inventing the Organizations of the 21st Century Initiative at the MIT Sloan School of Management and the Division of Research at the Harvard Business School.

work that needs to be performed. Under work pressure, service personnel struggle to keep a balance between the flows of incoming and outgoing orders while maintaining reasonable working hours and sustaining service quality. How a service organization responds to work pressure is a critical determinant of service quality, employee satisfaction, and the overall profitability of the service firm. Consider the following excerpts from interviews with nurses of a prestigious teaching hospital in the midst of consolidating functions after a recent merger:

> "It's not the desirable place it used to be to work. . . . You work very, very hard while you're here. . . . Nothing, absolutely nothing, laid back about this job. It's very high pressure, very high stress. . . . This is not going to be a place that I'm going to stay long term."
>
> "The nurses are feeling more rushed in getting patients through here. . . . I think patients are unhappy, and the nurses are unhappy because they're not able to give time and care to a patient that they wanted to. And I think probably things are being missed. I think patient care has suffered."[3]

Since services are produced and consumed instantaneously, service organizations are particularly vulnerable to these imbalances in supply and demand. That is, server and consumer have to be available at the same time for the service transaction to take place. This simultaneity leaves no room for a "finished goods inventory" to buffer the service delivery system from variations in demand. The problem of balancing supply and demand in services, however, is not simply a matter of absorbing short-term variations in customer orders; it persists for two reasons. First, growth results in an imbalance as investments to increase capacity struggle to keep pace with increasing demand. Over the last fifty years, the service sector has consistently been the fastest growing sector in the economy, and the situation is becoming critical as the U.S. economy is experiencing extremely tight labor markets. Nationwide unemployment stands at about 3.9 percent, and in some urban areas the figure is even lower. The U.S. Bureau of Labor Statistics estimates that, if current trends continue, by 2005 the total labor force will only be 2 percent higher than the total number of jobs in the U.S. economy.[4]

A second force sustaining systematic imbalances in supply and demand is the fact that service-sector productivity is improving more slowly than manufacturing productivity, resulting in increasing costs for the service sector.[5] Increasing operating costs translate into financial pressures that drive service organizations to process improvement and cost containment initiatives to seek efficiency gains. This continuous search for productivity gains might be the only way for service organizations to remain viable in a highly competitive environment. These initiatives, however, also ensure that the service organizations operate close to the balance point between supply and demand, thus increasing the risk of temporary imbalances or even systematic underinvestment in service capacity.

Since work pressure has a direct impact on service quality, employee satisfaction, and overall profitability, it is important to understand how a service organization responds to changes in work pressure and why it responds the way it does. More importantly, we must comprehend the consequences from each of these responses and identify what management can do about it.

Services Require an Integrative Approach

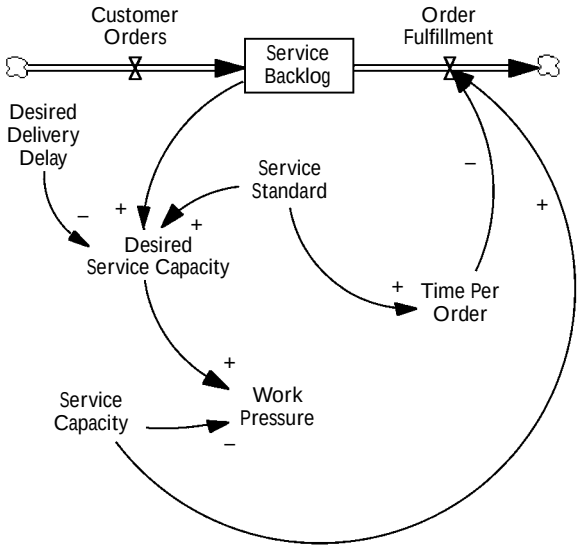
Services delivery processes differ from product manufacturing as the servers (employees) and the elements being processed (customers) are humans—with psychological attributes, perceptions, and expectations. Furthermore, services are produced in front of customers and often with direct collaboration from them, thus bringing employees and customers physically and psychologically close. Customers' perception of the service experience is not only affected by the conditions under which the service is delivered, but also by the employees' attitudes towards the customer. Similarly, employees' attitudes towards and perceptions of their job are influenced by customers' attitudes towards the service. This co-evolution of perceptions and expectations is further confounded by the fact that services are intangible, thus making it difficult to assess customer requirements and to fix an objective service standard. Clearly, the study of services requires an interdisciplinary approach; an integrated understanding of the organizational and behavioral components of the social systems that produce and consume the service, as well as the physical and technological characteristics of the service delivery system.

This research reported here uses the system dynamics method to explore different responses to work pressure. I created a formal simulation model that captures the structural characteristics of the service delivery process, management's and employees' decision-making processes, and the formation of expectations for customers and employees.[6] I found that the major recurring problems observed in service industry—erosion of service quality, high turnover, and low profitability[7]—can be explained by the organization's response to changes in work pressure. That is, the manner in which a service firm responds to work pressure determines whether the system will disappoint customers, employees, or shareholders. Furthermore, this research suggests ways to identify the structural characteristics that determine the preferred response in a given circumstance, thus making it possible to design high-leverage interventions and policies. Although quality, costs, and employee satisfaction are normally perceived as tradeoffs, I found that successful policies manage simultaneously to delight customers, employees, and shareholders.

Responses to Work Pressure

Work pressure in a service setting typically manifests itself through a backlog of customers waiting to be processed. The service backlog is the accumulated difference of the incoming customer orders and the order fulfillment rates.

FIGURE 1. Work Pressure: Imbalance in Supply and Demand

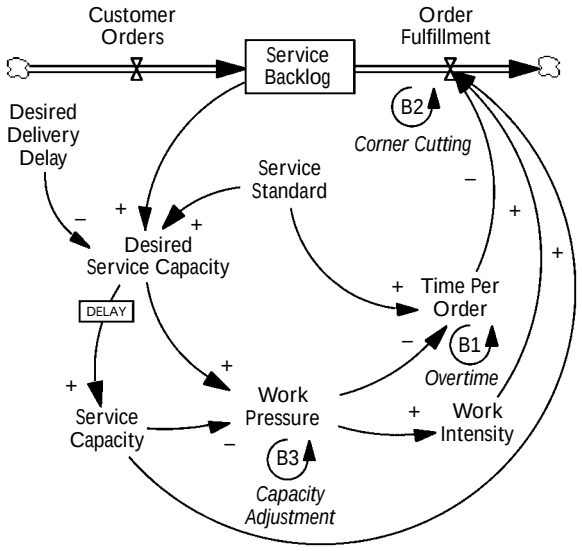


Arrows indicate the direction of causality. Signs ("+" and "–") indicate the polarity of relationships. A "+" means that, all else equal, if the cause increases (decreases), the effect increases above (decreases below) what it would otherwise have been. Similarly, a "–" means that, all else equal, if the cause increases (decreases), the effect decreases below (increases above) what it would otherwise have been.

Examples of customer backlog include the line in front of a cashier, callers on-hold in a calling center, or service requests waiting to be processed. From the service backlog, it is possible to determine the capacity required to ensure that service is provided within a certain time (delivery delay) and with a certain level of quality (service standard). Work pressure is the difference between the required capacity and the service capacity available (see Figure 1). If a target delivery delay and a service standard must be maintained, then the larger the service backlog, the higher the work pressure that the service providers feel.

The possible responses to work pressure in a service environment are limited. An option is to limit the flow of incoming orders. This can be done in different ways. The first possibility is to limit the holding capacity of the backlog. For example, if the bar in a restaurant has no more places available, then potential customers can infer that the waiting time for the service (dinner) is so long that it is not worth the wait. Similarly, some call centers report a busy signal for the caller once the number of calls waiting to be processed exceeds certain limit. A second alternative is to limit demand through price, or to use price incentives to balance the variations in customer demand. Once customers have entered the service setting, however, short of customers deciding to opt out of the waiting line (hanging up the phone or canceling their order), there are only three possible actions to reduce the backlog: Employees can work harder, employees can reduce the time allocated to each customer, or management can decide to increase service capacity.[8]

FIGURE 2. Responses to Work Pressure

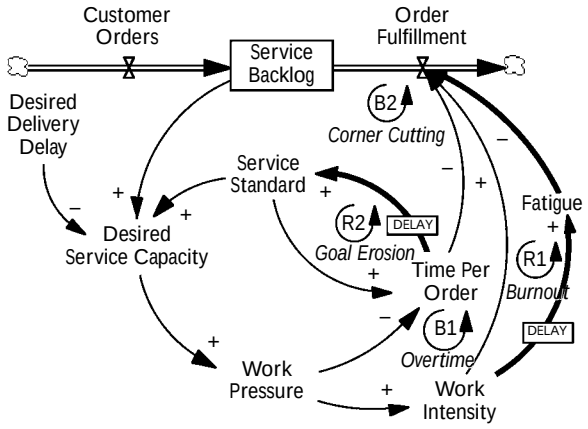


Delays separating the cause and the effect are indicated in the diagrams by delay boxes. Balancing loop polarity (denoted by B in the loop identifier) indicates a regulating (negative) feedback loop.

The first option for service providers experiencing work pressure is to increase their work intensity (WI), i.e., to work harder. This normally translates into providers' reducing the number and length of breaks that they take during the day. If this is not enough, service providers will eventually incur overtime. As all the other responses to work pressure, increasing the WI creates a balancing mechanism that attempts to maintain work pressure within certain limits. A higher work intensity increases the order fulfillment rate, thus reducing the service backlog and eliminating the work pressure (see the Overtime Loop in Figure 2). Under extreme work pressure, management might decide to reduce all training, planning, and improvement efforts and encourage service personnel to allocate all time available to order processing. This is typical behavior for organizations facing major deadlines.

A second response to changes in work pressure is for service providers to adjust the time per order (TPO), i.e., the time allocated to each customer order. Although the TPO is normally determined by the firm's service standard, in times of high work pressure service personnel can attempt to process each customer faster. Speeding up service transactions might be as simple as reducing the time spent in pleasantries with the customer, but it might also extend to eliminating post-service documentation or some core aspects of the service delivery process, i.e., "cutting corners" from the full-service transaction. Given a constant service capacity, reducing the time per order effectively increases the order fulfillment rate, thus reducing the existing backlog and eventually eliminating the work pressure (see the Corner-Cutting Loop in Figure 2).

FIGURE 3. Consequences of Employees' Responses to Work Pressure



Reinforcing loop polarity (denoted by R in the loop identifier) indicates a self-reinforcing (positive) feedback process.

The third response to changes in work pressure is to invest in service capacity (SC). Higher service capacity increases the order fulfillment rate, which in turn reduces the service backlog, the required service capacity, and work pressure (see the Capacity Adjustment Loop in Figure 2). Increasing service capacity can be accomplished by increasing the number of employees, investing in equipment and technology, or redesigning processes. Regardless of the selected alternative, adjustments of service capacity are not instantaneous and, in some cases, the introduction of new resources (computers or employees without experience) or the adoption of new technologies is disruptive in the short term.[9]

Performance Traps

Since increasing service capacity usually involves higher expenditures and longer delays than the other two alternatives, companies usually opt to allow their service personnel to absorb small variations in work pressure. Unfortunately, using employees' responses to work pressure (WI and TPO) over an extended period has some undesirable consequences. First, sustained periods of high WI cause fatigue that eventually reduces the productivity of the service providers. Reduced productivity of service providers leads to a lower order fulfillment rate that, everything else being equal, translates into a higher service backlog, further increasing work pressure and forcing service providers to work even harder (see Burnout Loop in Figure 3).[10] The second unintended consequence of using employees' responses to work pressure is subtler. Because of a lack of an objective service standard, extended periods of reduced TPO tend to modify the service standard as employees, managers, and customers adjust their service expectations to past performance.[11] With lower service standards, the time allocated per order is reduced, thus reducing the service standard even further (see Goal Erosion Loop in Figure 3). Note that these unintended

consequences are reinforcing traps (bad performance yields even worse results) that erode the work quality for employees or the overall service quality of the firm.

How should organization respond to changes in work pressure? The choice is not obvious. Only increasing service capacity avoids the performance traps of employee burnout or erosion of service standards. Adjustments of service capacity, however, are not instantaneous and directly affect the cost structure of the service firm. To a certain extent, it seems desirable to use the flexibility and immediacy of the employees' responses to deal with short variations in work pressure. While the individual effects of these responses to work pressure are simple to understand, it is more difficult to comprehend the tradeoffs presented by the interaction of these responses and to assess their long-term consequences. A well-calibrated and tested simulation model can help in understanding these dynamic tradeoffs.

Responses of a Lending Center

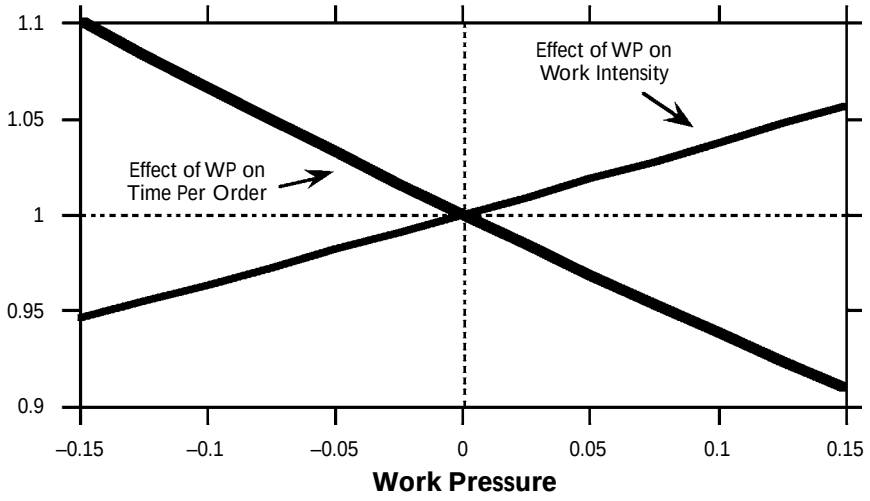
In 1990, a major retail bank in the UK sought to cut costs by moving back-office operations from branches to centralized processing centers in more affordable locations. Under this scheme, a Lending Center serves as a back-office for managing personal loans, credit cards, and small business accounts. Work arrives to the Lending Center by telephone (customer inquiries), mail (customer requests and communications with branches), and daily computer-generated reports identifying problematic accounts that require immediate action (such as overdrafts and missing payments). Most requests produce either a letter or a telephone conversation with the customer. Tasks are monitored against standard processing times, and there is a clear expectation that all tasks will be processed within 24 hours of their arrival.

While the strategy to centralize and standardize the back-office operations had improved the productivity of the lending officers and the quality of the lending book,[12] employees with direct customer contact had some concerns about the level of service provided. When asked how they felt the Lending Centers were working for the bank, here is what two lending officers had to say:

> "The feedback you get back [from the customer] is 'I'm dehumanized, I just became a number. I can no longer talk to you as a person, you just treat me as a number.' . . . we have lost the customers along the way."

> "I think it has been very effective in actually improving the lending book because of the monitoring system. Now that we've seen [the reduction of the risk index] through, because of the various pressures on us, we are going to be asked to be more proactive in selling. . . . We just don't have the relationship basis to sell effectively. The customers have said that they become a number; and in a way they have. . . . It is difficult to sell that way."

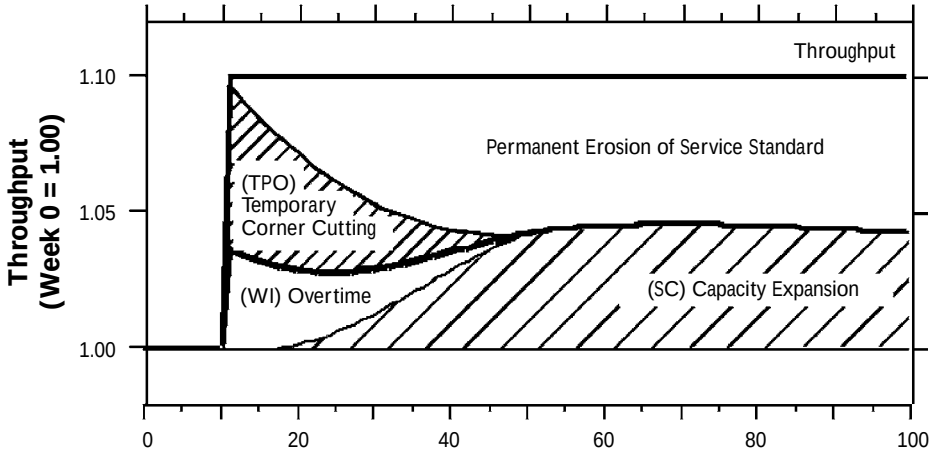
FIGURE 4. Lending Officers' Response to Work Pressure



Fieldwork was done at a Lending Center supporting 20 branches in the West End region of London. Data collected for the study included: time series for the operational metrics of the service center; interviews with employees, their managers, and staff inside and outside the Lending Center; direct observation; and archival data such as procedure manuals and training materials. I used these data to specify the decision rules of employees and managers in the model, as well as the physical constraints of the service delivery process. The adequacy of the model was tested by evaluating the rationality of the policies estimated for decision makers and the model's ability to replicate historical behavior for time per order, work intensity, order fulfillment, and multiple metrics of service capacity.

Analyzing the operational performance, I found that, under pressure to increase output, lending officers are much more willing to cut corners (reduce TPO) and only reluctantly work longer hours (increase WI). Figure 4 shows the lending officers' estimated response to work pressure. In the absence of work pressure (WP=0), lending officers work their regular hours and allocate to each customer order the effort dictated by the service standard—effect of work pressure on WI and TPO is equal to one (neutral). If work pressure rises above its neutral level (0), employees simultaneously increase their work intensity and reduce the time per order. Their response on cutting corners, however, is almost twice as aggressive as their increase of work intensity. Although in interviews and surveys employees claimed a deep concern for the "standard of customer service" of the 15 loan officers interviewed, all but one admitted to reducing their effort to document transactions and sell additional products in times of high work pressure. Note that consistent with the emphasis the monitoring

FIGURE 5. Stack Responses to a 10% Increase in Demand



system places on processing customer orders the same day they arrive, employees' response to work pressure is at all times enough to compensate for the change in work pressure. For example, a 5 percent increase in work pressure is handled through a 1.8 percent increase in work intensity and a 3.2 percent reduction of time per order.

Consistent with the interviews with loan officers, simulations showed a 4 percent reduction in the service standard—measured in time per order—for the period when data were available. The erosion persisted even when the model was initiated in equilibrium, ruling out transient effects from the opening of the Lending Center. Furthermore, extended models simulations showed that the erosion of service continued well beyond the point where service capacity and service demand had reached equilibrium (no work pressure).

To isolate how the interactions among the three responses to work pressure—increasing WI, reducing TPO, and increasing SC—generate the erosion of the service standard, I tested the model without the random variations of customer orders. Initially, the service center is balanced: employees are not working overtime and customers receive the expected level of service. At week 10, the simulation introduced and sustained, a 10 percent increase in customer orders. Since the Lending Center has a commitment to process all orders within 24 hours (i.e., constant delivery delay), the excess orders require the Lending Center to increase its throughput by 10 percent. Figure 5 shows the contribution to throughput achieved by each of the responses to work pressure, as well as the erosion of the service standard. For example, five weeks after the introduction of the incremental demand (week 15), service capacity has not yet been expanded, and the service standard has already eroded 2 percent because of aggressive corner cutting during the initial weeks of work pressure. The remainder of the

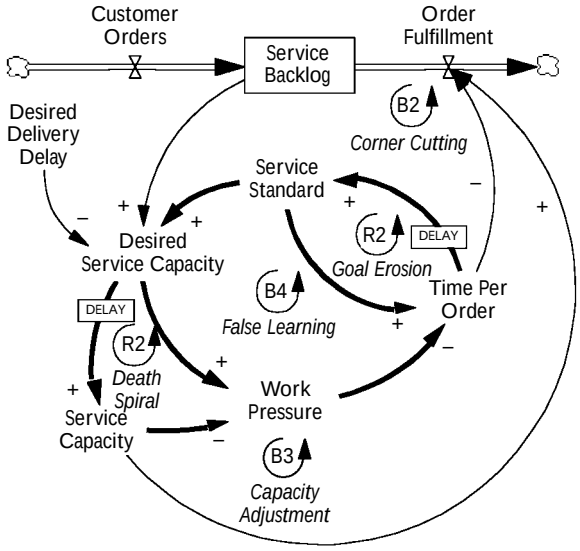
required increase in throughput is achieved through a 3 percent increase in overtime, and a 5 percent reduction of time per order beyond the current standard.

The combination of responses is effective in immediately increasing throughput—at all times the combination of responses adds up to the required 10 percent increase. However, the timing and strength of these responses differ substantially. As explained above, the initial reduction of TPO is almost twice as aggressive as the increase in WI. Furthermore, whereas employees' responses to work pressure (TPO and WI) are essentially instantaneous, the adjustment of service capacity is slow, peaking after 25 weeks. There are several reasons for the lag in the response to adjust SC. First, although performance metrics are available on a weekly basis, these metrics are summarized and analyzed only at the end of the month. Next, to smooth out the high frequency variations in customer orders, managers adjust their estimate of required service capacity with an average lag of 4 months. The delay in adjusting authorized labor achieves its purpose of filtering out variations in customer orders and is consistent with management's imperative to control costs. Once labor has been authorized, it takes on average seven months for the hiring process to bring a new employee into the Lending Center. Finally, rookies are on average 35 percent as productive as experienced personnel, with an average delay of 12 months before becoming fully productive. The combination of cautious hiring policies, hiring delays, and long training requirements cause service capacity to react slowly to changes in demand. The consequence of the slow adjustment of SC is that temporary variations in work pressure must be accommodated through either cutting corners or working overtime.

Model simulations with historical variation patterns showed that, as expected, employees absorb small increases in work pressure arising from variations in demand and absenteeism by reducing TPO and increasing WI. Although reducing TPO enabled an immediate increase in throughput, it also resulted in the erosion of the internal service standard (the Goal Erosion Loop in Figure 6). Furthermore, in the absence of direct, reliable, and trusted measurements of customer satisfaction, management interprets the reduction in TPO and service standard as productivity gains, and adjusts the required service capacity accordingly. The adjustment of desired service capacity through the erosion of the service standard should eventually eliminate work pressure and bring the system to equilibrium (the False Learning Loop in Figure 6).[13] However, the reduction of desired service capacity also translates, with a delay, into a reduction of the actual service capacity. Lower service capacity further increases work pressure on the service delivery personnel, who in turn reduce TPO, thus locking the system into a vicious cycle of eroding standards and diminishing service capacity (the Death Spiral Loop in Figure 6).

The relative strength and timing of the responses (TPO>WI>SC) accounts for the observed erosion of service standards. By the time hiring reacts to the changes in customer orders and new employees are trained, the required service

FIGURE 6. Erosion of Service Standard



capacity has eroded with the service standard. Regressions showed that the reduction of the service standard experienced during the period for which data were available translated into a 50 percent reduction in expected sales. As large as lost sales can be, they still underestimate the hidden costs of a low service standard, since high work pressure also translates into errors in documentation and higher rework rates.

Policy Analysis

Once the reasons for quality erosion were understood, I used the model to develop policy recommendations. Specifically, we were interested in developing policies to maintain service quality without compromising the organization's ability to respond to demand fluctuations and without carrying excess capacity. There are only three ways in which the Lending Center's response profile for changes in work pressure (Figure 5) can be modified while still maintaining full responsiveness: expediting the adjustment of service capacity, reducing the effect of work pressure on time per order, or reducing the rate at which the service standard erodes. These three changes correspond to moving "up" each of the lines separating the responses in Figure 5. Following is a list of possible implementations of these strategies.

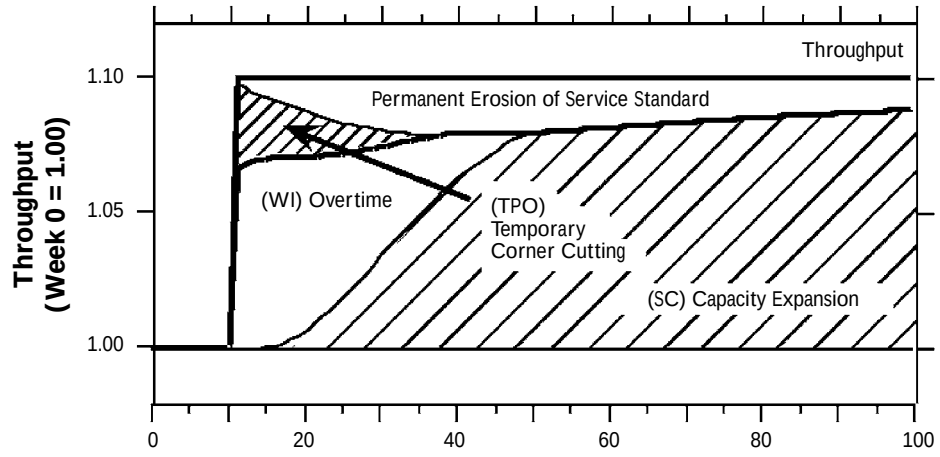
- *Expediting the Adjustment of Capacity*—Since the erosion of the internal service standard occurs when work pressure is high, one obvious policy is to ensure that service capacity is acquired before the standard can erode. Capacity expansion can be expedited by reducing the delays governing

the capacity adjustment process and creating a more responsive hiring process. A second option is to reduce the time it takes rookies to become fully productive. Accelerated learning rates can be achieved either by hiring employees with greater initial effectiveness or by increasing the intensity of their training programs. Unfortunately, the option to accelerate learning curves is rarely available in high-contact services that require job-specific knowledge.

- *Reducing the Effect of Work Pressure on Time Per Order*—The positive feedback driving the erosion of the service standard is triggered by cuts in time per order caused by high work pressure. Reducing employees' willingness to cut corners should weaken the Goal Erosion Loop and slow the decline of the service standard. This normally requires management to become aware of the implications of poor service—lost sales, rework, and customer defections—and to persuade employees to avoiding them; even at the cost of slower transactions. Of course, if the time spent with customers were completely unaffected by work pressure, there could be no quality erosion. Such a rigid policy, however, is unfeasible in high-contact services where employees have considerable autonomy in deciding how they respond to each customer. A more realistic policy should aim to distribute employee responses to work pressure evenly between corner cutting and overtime. This can be done by reducing the flexibility of the service encounter (through process standardization and documentation) or by increasing the relative attractiveness of overtime (by creating high empathy with customers or increasing overtime compensation).
- *Reducing the Erosion Rate for the Service Standard*—The intent of the previous two policies is to reduce the usage of TPO as a coping strategy for changes in work pressure. Under unavoidable variations of work pressure, however, maintaining service quality requires explicit pressure to prevent the erosion of the standard. An external norm for service quality may provide sufficient counterpressure to halt standard erosion. In some industries, such external norms are developed as part of the professional training of service providers. More often, management must take an active role in forming the service standard by articulating clear and consistent expectations for service quality and monitoring performance against them.

How much should each of these levers be pushed in order to maintain service quality and full responsiveness to variations in demand? Two separate analyses are required to answer this question. First, it is necessary to determine how successful a policy is in achieving its stated goal. Because of non-linear relationships and multiple feedback mechanisms interacting in a service setting, it is difficult to predict the effectiveness of a proposed policy and the unintended consequences that it might generate. Experimentation with a computer simulation model is an effective way to assess the impact of each proposed policy and its interactions with other elements of the system. For example, despite the apparent benefits of reducing the time for employees to become fully trained,

FIGURE 7. Stack Responses to a 10% Increase in Demand after Implementation of Policy Recommendations



I found that this policy had limited impact in the Lending Center since on average only 6 percent of its employees are inexperienced at any given time.

Once a policy is deemed effective, it is necessary to assess how easy it would be to implement the policy in the real system. Policy changes require structural modifications—i.e., changes in norms, processes, and decision making. The feasibility of structural changes in a particular situation is a function of the existing conditions and the cost to implement the changes. For example, in the case of the Lending Center it was desirable to reduce the delays of the labor authorization and hiring processes—a combined delay of 11 months. However, the existing set of relationships between the Lending Center and the district office of human resources made it feasible to think only of a 50 percent reduction in these delays.

In the case of the Lending Center, no one single policy was successful in stopping the erosion of the quality standards and a combination of the three policies described above was recommended. Figure 7 shows the modified response profile for the Lending Center after the three policies were implemented simultaneously. Notice that WI takes a much larger share of the required throughput, as quality norms prevent the extended use of TPO. Faster capacity acquisition and more emphasis on overtime did reduce the effect of work pressure on time per order and the service standard. However, in order to raise the standard after its initial erosion, it was necessary to introduce managerial pressure that would drive employees to deliver at levels above their current internal standard. In general, I found that policies that maintain a balance among the different responses are more successful at maintaining throughput and avoiding the performance traps. That is, if the excess work pressure is distributed among

multiple responses, the unintended consequences from all the responses are minimized.

Structural Determinants of Response Flexibility

The relative intensity of the responses to work pressure in the Lending Center are determined by the structural characteristics of its service delivery system, specifically the need to customize service transactions and the delays in hiring and developing employee skills. Customization inhibits standardization of the service delivery process, thus allowing service employees to reduce service scope in response to work pressure. A significant but slow learning curve reduces the speed at which service capacity can be acquired. The specifics certainly vary from industry to industry—for example, service settings with high professional standards will have stronger quality pressure and slower erosion of standards. However, given the broad prevalence of on-the-job training, delays in capacity acquisition, and the intangibility of services, the structure that can lead to erosion of service quality is likely to be common throughout the service sector.

Erosion of service quality, however, is not the only possible outcome from work pressure. Other service settings have a different set of dominant structural characteristics that affect the strength of the responses to work pressure. For example, the use of the reduction of TPO as a way to deal with work pressure is limited if the service delivery process has been standardized, or if professional standards constrain the customer-facing personnel to provide a certain service level. The WI response is limited by the working hours in the service setting and the willingness and incentives that employees have for working overtime. Finally, the responsiveness of changes in SC is limited by the amount of training required, the speed at which additional service capacity can be acquired, and the managerial policies in place. A detailed list of the main constraints on each of the response mechanisms is presented in Table 1. Each of the limiting factors relates to a structural characteristic of the service setting.

Once the structural characteristics of the service setting are mapped into the basic responses to work pressure, it is possible to identify the relative strength of the responses to work pressure that could be expected from those structural components. For example, in a service with a standardized (rigid) delivery process (low TPO flexibility) and relative short training requirements (high SC flexibility)—e.g., a fast-food restaurant—the relative strength of the responses to work pressure that could be expected is: first, increase of WI, then increase SC, and, probably very weakly, a reduction in TPO. Table 2 presents examples of service settings with stylized structural characteristics and their expected response preference to changes in work pressure. It is worth noting that services with different structural characteristics could show the same ranking of responses. For example, a capital-intensive firm with fixed standards of service—e.g., a utility—that is forced to use equipment beyond the natural

TABLE 1. Factors Limiting the Flexibility of Responses to Work Pressure

Response	Factors Limiting Flexibility of Response
Time Per Order	• Standardized Service Delivery Process (Low Customization) • Professional Quality Standard (High Customization) • Quality Sensitive Customers • Good Information of Quality Performance
Work Intensity	• Constraints on Working Hours • High Customer Contact Time • Regulated Work-Hours (Airline Pilots) • Employee's Lack of Empathy with Customers
Service Capacity	• Capital Intensity • Long Training Requirements • Professional Certification • Long Hiring Delay • Financial Pressures

maintenance cycle to satisfy an increase of demand would have the same ranking of responses as the fast-food restaurant (WI>SC>TPO).

Because the similar response rankings create similar dynamic behaviors, regardless of the structural limitations causing the ranking, the relative strength of the responses to work pressure can be used as a way to classify the dynamics that characterize services. Such classification allows for a reduction of the dimensional characteristics needed to differentiate service settings and creates a direct linkage between the dominant structural characteristics of the service setting, its behavior, and potential policy recommendations.

For instance, the high turnover observed in entry-level service jobs is not surprising when exploring the structural characteristics of the settings where it is normally encountered. Jobs requiring few skills are characterized by standardized service delivery process (low TPO flexibility), low face-to-face contact (high WI flexibility), and very low margins that discourage the investment of service capacity (low SC flexibility).[14] These structural characteristics suggest an overwhelming preference of WI over the other potential responses to work pressure (WI>>SC>TPO). By limiting the other two responses, through tight standardization and austere capacity policies, management has—perhaps unintentionally—selected to disappoint employees and use them as the escape valve for changes in work pressure. It is, then, not surprising that employees would decide to leave this type of job.

Similarly, the erosion of service quality and burnout reported by the nurses in the teaching hospital (as described earlier) can be explained by the dominant structural characteristics of the health care delivery system and the responses that they generate. The attributes of health care practitioners, strict professional quality standards and high empathy with customers, combined with long training requirements and intensive capital investments to increase

TABLE 2. Response Flexibility: Examples, Structural Characteristics, and Expected Response Preference

Low TPO Flexibility

SC Flexibility	WI Flexibility: Low	WI Flexibility: High
High	**Ticket Sales** • Standardized delivery Process • Limited Work Hours • Low Training Requirements **SC>TPO>WI**	**Fast-Food Restaurant** • Standardized Delivery Process • Low Face-to-Face Contact • Low Training Requirements **WI>SC>TPO**
Low	**Airline Pilots** • Standardized Delivery Process • Regulated Work Hours • Long Training Requirements **SC>WI&TPO**	**Health Care** • Strict Professional Standards • High Empathy with Customers • Long Training/Capital Intensive **WI>TPO>SC**

High TPO Flexibility

SC Flexibility	WI Flexibility: Low	WI Flexibility: High
High	*	*
Low	**Claims Adjusting Process** • High Customization • Limited Work Hours • Long Training Requirements **TPO>WI>SC**	**Maintenance Crew** • High Customization • Low Face-to-Face Contact • Long Training Requirements **TPO&WI>SC**

* The combination of high SC flexibility with high TPO flexibility is not feasible since high customization (TPO flexibility) implies long training requirements (low SC flexibility).

capacity, yield a response profile that would first tap into WI to compensate work pressure. If high WI is sustained, health care practitioners will eventually reduce the TPO and erode their service standard. This predicted response is consistent with the employee burnout and erosion of the service standard reported throughout the health care industry.

There are other reasons why a service firm might exhibit an undesired pattern of behavior—for instance, high turnover might be driven by a tight labor market. Although those environmental issues can be easily incorporated into the analysis, the model focuses on endogenous explanations for the observed behavior, i.e., most variables are internal to the service center and under managerial control. The endogenous perspective, by making the tradeoffs among options explicit, allows the model to be used to explore alternative intervention strategies for improving performance. Finding an optimal mix of policies to achieve a target quality level and responsiveness to demand still requires the specification of a detailed cost function associated with each response and policy lever. However, by establishing the link between structural characteristics of the service setting, flexibility of responses to work pressure, and undesired behaviors, the framework presents a full description on how the feedback structure of a system generates its behavior. Based on this framework, it is possible to design structural changes (high-leverage policies) to avoid the undesired consequences of work pressure. Undesired behavior (high turnover, erosion of service quality, low profitability) is, in most cases, a symptom of an unbalanced response to changes in work pressure (work intensity, cutting corners, excess capacity). Balancing the responses to work pressure should be a matter of decreasing the flexibility of the abused response and increasing the flexibility of the untapped responses. A direct understanding of the structural characteristics that determine the flexibility of the each response makes the diagnosis and selection of leverage points easier.

Notes

1. B. Tuchman, "The Decline of Quality," *New York Times Sunday Magazine*, November 2, 1980, p. 38; J. Main, "Toward Service Without a Snarl," *Fortune*, March 23, 1981, p. 58; M.R. Feinberg and A. Levenstein, "It's Not My Job, Man," *Wall Street Journal*, November 11, 1985; S. Koepp, "Why Is Service So Bad? Pul-eeze! Will Somebody Help Me?" *Time*, February 2, 1987, p. 46; D. Brady, "Why Service Stinks," *Business Week*, October 23, 2000, pp. 118-128.
2. "American Customer Satisfaction Index," American Society for Quality, 2000, <http://acsi.asq.org/>.
3. D.B. Weinberg, "Why Are the Nurses Crying? Restructuring, Power, and Control in an American Hospital," unpublished Ph.D. dissertation, Harvard University, Cambridge, MA, 2000.
4. D.S. Friedman, "Help Wanted," *The McKinsey Quarterly*, 1 (1998): 34-44.
5. W.Baumol, S.B. Blackman, and E. Wolf, *Productivity and American Leadership* (Cambridge, MA: MIT Press, 1991).
6. For a full description and documentation of the model, see R. Oliva and J.D. Sterman, "Cutting Corners and Working Overtime: Quality Erosion in the Service Industry," *Management Science*, 47/7 (2001): 894-914.
7. For evidence on erosion of service quality, see D. Brady, "Why Service Stinks," *Business Week*, October 23, 2000, pp. 118-128; Oliva and Sterman, op cit. For erosion of job scope and employee satisfaction, see L. Schlesinger and J. Heskett, "Breaking the Cycle of Failure in Services," *Sloan Management Review*, 32/3 (Spring 1991): 17-28. For low profitability in the service sector, see W.J. Baumol, S.A.B.

Blackman, and E.N. Wolff, "Unbalanced Growth Revisited: Asymptotic Stagnancy and New Evidence," *American Economic Review*, 75/4 (1985): 806-817; P.T. Harker, "Introduction: Service-Sector Productivity—The MS/OR Challenge," *Interfaces*, 25/3 (1995): 1-5.

8. Note that prioritizing customer orders—or subjecting them to a triage process—does not help reduce work pressure. Through prioritization, it is possible to speed up the response to an important customer or to a person requiring the service more promptly, but the overall load over the service delivery system remains unchanged.
9. For a discussion on the impact of improvement efforts in the order fulfillment rate, see N.P. Repenning, L.J. Black, and P. Gonçalves, "Why Good Process Sometimes Produce Bad Results," published in this issue [*California Management Review*, 43/4 (Summer 2001)]; E. Keating, R. Oliva, N.P. Repenning, S. Rockart, and J.D. Sterman, "Overcoming the Improvement Paradox," *European Management Journal*, 17/2 (1999): 120-134.
10. For a survey of evidence on the effects of overtime on labor productivity, see H.R. Thomas, "Effects of Scheduled Overtime on Labor Productivity: A Literature Review and Analysis," PTI9107, The Pennsylvania Transportation Institute, Pennsylvania State University, University Park, PA, 1990. Homer explores the burnout mechanisms from a systems dynamics perspective. J.B. Homer, "Worker Burnout: A Dynamic Model with Implications for Prevention and Control," *System Dynamics Review*, 1/1 (1985): 42-62.
11. See J. March and H. Simon, *Organizations* (New York, NY: Wiley, 1958); R. Cyert and J. March, *A Behavioral Theory of the Firm* (Englewood Cliffs, NJ: Prentice Hall, 1963). For empirical evidence, see T.K. Lant, "Aspiration Level Adaptation: An Empirical Exploration," *Management Science*, 38/5 (1992): 623-644.
12. Quality of the lending book is measured through a risk index calculated by weighting the each account's risk grade by the loan amount. The lending centers had yielded significant improvements of the book's risk index through centralization and standardization of the lending criteria.
13. Notice in Figure 5 how the employee's responses disappear after the permanent erosion of the service standard and service capacity reach an equilibrium and work pressure returns to zero. The slight adjustment between capacity expansion and the permanent erosion of service standard after week 48 is due to an overshoot of the hiring policy that results in negative work intensity. The system slowly reaches equilibrium (through employee attrition) at the service level attained when work pressure first reached equilibrium. In this particular test, the simulated organization increased its throughput 10 percent by reducing the internal standard of customer service 5.4 percent and increasing service capacity 4.1 percent. Throughput (orders/week) is defined as service capacity (man-hours/week), divided by time per order (man-hours/order). (1+0.041)/(1-0.054)=1.1.
14. For an account of an employee attempting to survive on low satisfaction, high turnover services, see B. Ehrenreich, *Nickel and Dimed: On (Not) Getting by in America* (New York, NY: Metropolitan Books, 2001). See also L. Schlesinger and J. Heskett, "Breaking the Cycle of Failure in Services," *Sloan Management Review*, 32/3 (Spring 1991): 17-28.

Back to the Top