Oliva, R. and J. Sterman (2010). Death Spirals and Virtuous Cycles: Human Resource Dynamics in Knowledge-Based Services. *Handbook of Service Science*. P. P. Maglio, C. A. Kieliszewski and J. C. Spohrer (eds). New York, Springer: 321-358.

Death Spirals and Virtuous Cycles

Human Resource Dynamics in Knowledge-Based Services

Rogelio Oliva

Mays Business School

Texas A&M University

John D. Sterman

Sloan School of Management

Massachusetts Institute of Technology

While the productivity and quality of manufactured products steadily improve, service sector productivity lags and quality has fallen. Many service organizations fall into "death spirals" in which pressure to boost throughput and control costs leads to worker burnout and corner cutting, lowering service quality, raising costs while revenue falls, forcing still greater cuts in capacity and even lower quality. We present a formal model to explore the dynamics of service delivery and quality, focusing on the service quality death spiral and how it can be overcome. We use the system dynamics modeling method as it is well suited to dynamic environments in which human behavior interacts with the physics of an operation, and in which there are multiple feedbacks connecting servers, managers, customers, and other actors. Through simulations we demonstrate that major recurring problems in the service industry—erosion of service quality, high turnover, and low profitability—can be explained by the organization's internal responses to work pressure. Although the reinforcing feedbacks can operate as virtuous as well as vicious cycles, the system is biased toward quality erosion by basic asymmetries and nonlinearities. We show how, with the right mix of policies, these same feedbacks can become virtuous cycles that lead to higher employee, customer satisfaction and additional resources to invest in still greater service quality improvement.

Introduction

Increasing class size and teacher burnout, shorter hospital stays and longer waits for emergency care, long waits on hold and unsatisfying conversations with customer service agents—these are all symptoms of poor quality in knowledge-based services. Despite the growing importance of services and service quality as sources of competitive advantage, the quality of service delivery in the United States is not improving and in many cases is falling. Complaints about poor service are staples in the popular press and online (see, for example, Aho, 2008; McGregor et al., 2009). While the quality of most manufactured products has improved over the past few decades, the American Customer Service Index fell to 72.5 in 2008, down 4% from its 1995 level (see http://www.theacsi.org). What explains the divergence? Why has quality improved so much for products but fallen on average for services? Here we develop an integrated dynamic model to explore the sources of persistent low service quality.

Services differ from manufacturing because they cannot be inventoried, so balancing capacity and demand is more difficult than in manufacturing. More important, services differ from manufacturing because they are produced in the context of a personal interaction between the customer and the server. Services are often produced in front of customers and often in direct collaboration with them, thus bringing employees and customers physically and psychologically close. The quality of a service interaction is necessarily a subjective judgment made by the individual customer. Feelings and emotions matter. Because customers have different backgrounds, knowledge, needs, and expectations, services are harder to standardize than manufacturing. Perceptions of procedural fairness and respect are important. Customers do not evaluate service quality solely in terms of the outcome of the interaction (e.g., did the doctor correctly diagnose my illness?) but also consider the process of service delivery (e.g., did the doctor take the time to hear me out, listen with empathy and treat me with respect—or rush through the appointment as quickly as possible to get to the next patient?).

Customers' perceptions of the service experience are not only affected by the conditions under which the service is delivered, but also by employee attitudes towards the customer. Similarly, employees' attitudes towards and perceptions of their job are influenced by customers' attitudes and behavior. The 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.

The study of services therefore requires an interdisciplinary approach that integrates the physical and technological characteristics of service delivery with the organizational and behavioral features of the social systems in which service delivery is embedded. Such interdisciplinary studies are coming to be known as "services sciences" (Chesbrough, 2005; Chesbrough and Spohrer, 2006; Horn, 2005; Maglio et al., 2006). Here we use the system dynamics modeling method (Sterman 2000) because it is well suited to dynamic environments in which human

behavior interacts with the physics of an operation, and in which there are multiple feedbacks connecting servers, managers, customers, and other actors.

We seek to understand the persistence of capacity problems in services and the persistent failure of service quality to improve over the past few decades. The tools of process improvement and the quality movement have been applied to service delivery just as they have to manufacturing, yet the quality gap continues to widen. Why? Since services are produced and consumed simultaneously, with no finished goods inventory, service providers are particularly vulnerable to imbalances between supply and demand. The problem of balancing supply and demand in services, however, is not simply a matter of absorbing short-term variations in customer orders. Rather, chronic undercapacity persists for two reasons. First, organizations facing growing service demands struggle to acquire capacity fast enough. Over the last fifty years, the service sector has consistently been the fastest growing in the economy. Second, service-sector productivity improves slowly compared to manufacturing (Baumol et al., 1991). Technological progress has dramatically increased output per person in manufacturing, but one taxi driver is still needed per taxi. Low productivity growth drives service organizations to seek efficiency gains and impose cost containment initiatives. The continuous pressure to "do more with less" pushes service organizations to operate with little margin to accommodate demand variability. We show that such policies not only lead to poor quality when demand temporarily rises, but can trigger a set of self-reinforcing processes that lead to the persistent, continual erosion of service quality, service capacity, and the customer base. These positive feedbacks operate as vicious cycles that can drag an organization into a death spiral of declining quality, customer loss, budget cuts, higher work pressure, poor morale, higher employee attrition, and still lower quality. Poor service can destroy a firm's brand and erode sales. In contrast, high quality service boosts customer loyalty, repeat business and favorable word of mouth that can increase growth and market share.

Here we present a formal model to explore the dynamics of service delivery and quality, focusing on the service quality death spiral and how it can be overcome. The work builds on foundations presented elsewhere (e.g., Oliva, 2001; Oliva and Bean, 2008; Oliva and Sterman, 2001; Sterman, 2000, Chapters 12 and 14), but adds additional structure we have identified over the last ten years working with organizations that provide knowledge-based services. We present the model iteratively, beginning with the dynamics of human capital. We then add additional structures, including the interactions of employees and customers, the workweek, standards for customer service, hiring and training, customer responses to service quality, and budgeting. We explore how the dynamics change as the model boundary expands. A documented version of the model is available for experimentation under different assumptions.[1]

The paper is structured as follows. We first present a structure that captures the dynamics of the experience learning curve. We then introduce notion of work pressure, the gap between required and available service capacity, and explore the service providers respond to imbalances. We next expand the model to include the

[1] http://iops.tamu.edu/faculty/roliva/research/service/handbook/.

feedback effects of performance on the market and we then add budgeting and financial constraints on hiring. We close with policy recommendations.

Service capacity

We begin with the service organization's human capital, including hiring, training, and learning-by-doing. Learning-by-doing is well documented in diverse settings, including services (Argote and Epple, 1990; Darr et al., 1995). The importance of customization suggests potential for significant learning in high-contact service settings. When services involve personal and customized interaction between individual servers and customers, much of the learning gained through experience is embodied in the skills and behaviors of the individual workers.

Experience chains and learning curve

We model the individual learning curve of new employees as an "experience chain" (Jarmain, 1963). The workforce is divided into two populations: experienced and recently hired "rookie" employees (Figure 1). New hires are less productive than experienced employees, but gradually gain skills through experience, on-the-job coaching and mentoring. The effective workforce, measured in fully trained equivalent employees, is given by

Effective Workforce = Experienced Employees + Rookie Productivity Fraction * Rookie Employees. (1)

The stocks of rookie and experienced employees accumulate their respective flows of hiring, assimilation, and quits.[2] The model is initialized in equilibrium to facilitate testing.

Rookie Employees = INTEGRAL(Rookie Hire Rate – Rookie Quit Rate – Assimilation Rate, Initial Workforce * $\text{Rookie Fraction}_{ss}$) (2)

---

[2] We present the model with a minimum of mathematical notation, using "friendly algebra" in which the variables are named for the concepts they represent and correspond exactly to the simulation model. The INTEGRAL function denotes an accumulation, specifically:

Stock = INTEGRAL(Inflow –Outflow, Initial Stock)

is equivalent to

$$Stock_T = \int_{t_0}^{T} (Inflow - Outflow)dt + Stock_{t_0}$$

.

Experienced Employees = INTEGRAL(Assimilation Rate – Experienced Quit Rate,
Initial Workforce * (1 – Rookie Fraction$_{ss}$)) (3)

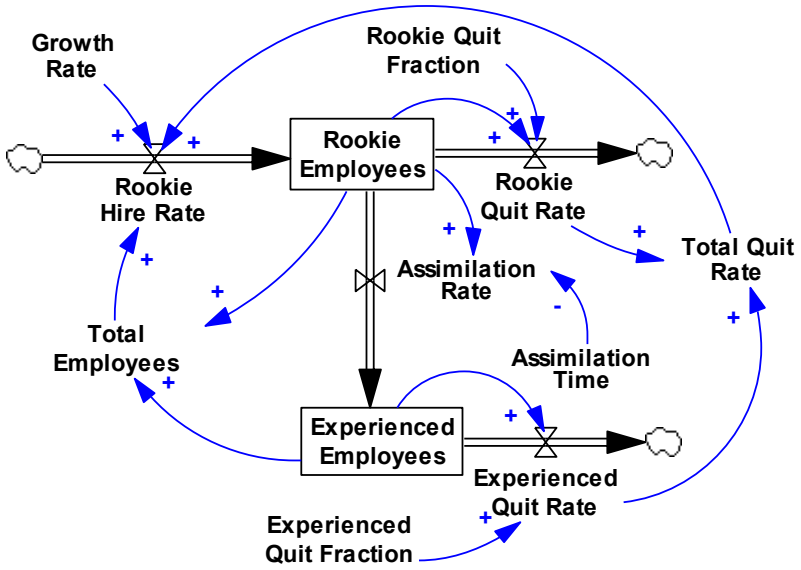


Figure 1. Experience chain structure.

where Rookie Fraction$_{ss}$ is the equilibrium fraction of rookies (eq. 13). Formulating the flows as first-order processes yields

Rookie Quit Rate = Rookie Employees * Rookie Quit Fraction (4)

Experienced Quit Rate = Experienced Employees * Experienced Quit Fraction (5)

Assimilation Rate = Rookie Employees / Assimilation Time. (6)

Average worker productivity is:

Average Productivity = Effective Workforce / Total Employees. (7)

Two parameters determine the speed and strength of the learning curve: the rookie productivity fraction—the productivity of rookies relative to fully trained employees—and the assimilation time—how long it takes rookies to become fully experienced. Figure 2 shows two simulations of the learning process. The solid line represents a setting where the service tasks are not difficult to master (e.g., a fast food restaurant). New hires have an initial productivity equal to 80% of a fully trained employee and an assimilation time of three months. The hashed line represents a more complex setting (e.g., financial services) where rookie productivity is only 20% of experienced employees and it takes an average of one year to become fully productive.

Finally, the total quit rate is the sum of quits from each employee cohort, total employees sums the stock of employees in each experience-level cohort, and the rookie fraction is the ratio of rookies to total employees:

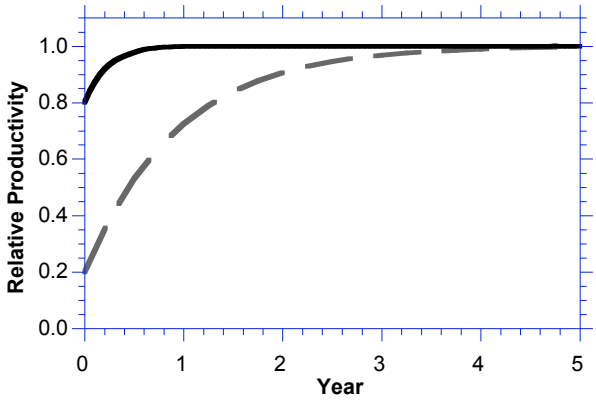


Figure 2. Examples of learning curves.

Total Quit Rate = Rookie Quit Rate + Experienced Quit Rate (8)

Total Employees = Rookie Employees + Experienced Employees (9)

Rookie Fraction = Rookie Employees / Total Employees. (10)

For purposes of testing, assume for now that the workforce grows at a constant exponential rate. That is, the firm replaces all those who quit and adds a fraction of the current total workforce:

Rookie Hire Rate = Total Quit Rate + Growth Rate * Total Employees. (11)

Below we replace this simple hiring formulation with an endogenous hiring rule that accounts for the adequacy of service capacity.

Despite its simplicity, the structure above illustrates fundamental dynamics of human resources. Figure 3 shows the effect of higher employee turnover on productivity. We assume a rookie productivity fraction of 30% and an assimilation time of 1 year. The simulation begins in equilibrium with headcount growth of 20%/year.

Figure 3a assumes an annual turnover rate of 20% while 3b assumes 80% annual turnover. In both cases the firm grows to 270 employees after five years. However, the employee mix in the two scenarios is quite different — the rookie fraction increases from 29% in the base case to 50% with high turnover. Growth causes significant experience dilution. Oliva et al. (2002) show how this structure caused service quality problems for a rapidly growing airline.

To understand the effects of the parameters on productivity, consider the rookie fraction when the system reaches steady state, i.e., when, the ratio of rookies to

experienced workers becomes constant (total labor might still be growing). Average productivity can be expressed as:

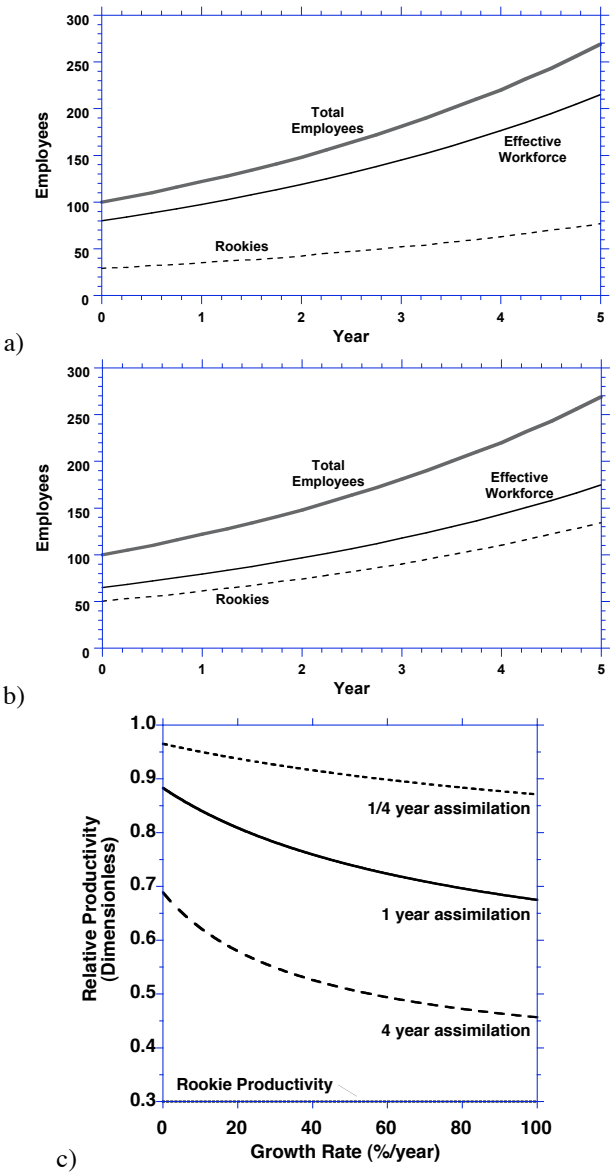


Figure 3. Effect of turnover on workforce productivity

Average Productivity =
(1 – Rookie Fraction) + Rookie Productivity Fraction * Rookie Fraction. (12)

The steady state rookie fraction is easily shown to be:

$$\text{Rookie Fraction}_{ss} = \text{Assimilation time}*(\text{Experienced Quit Fraction} + \text{Growth Rate})/(1 + \text{Assimilation Time} * (\text{Experienced Quit Fraction} + \text{Growth Rate})). \quad (13)$$

Figure 3c shows steady-state average productivity as a function of headcount growth for three different assimilation times, assuming rookie productivity is 30% and turnover is 20%/year. Slower assimilation or faster growth increases the steady-state rookie fraction, reducing average productivity. Without growth, average productivity is 97%, 88% and 69% of the fully experienced level with assimilation times of one-quarter, one, and four years, respectively, but drops to 91%, 62%, and 51% when headcount grows at 50%/year.

Mentoring

Rookies typically learn with the help and mentoring of experienced employees. On-the-job training, however, is not free. Mentoring reduces the time experienced personnel can allocate to their own work as they supervise rookies, demonstrate proper procedure and answer their questions. The effective workforce is thus determined by the effective number of experienced employees, which is the number of experienced workers net of the time they devote to mentoring:

Effective Workforce = Effective Experienced Employees +
Rookie Productivity Fraction * Rookie Employees (14)

Effective Experienced Employees = MAX (0, Experienced Employees –
Rookies * Fraction of Experienced Time Required for Training). (15)

Effective experienced employees are constrained to be nonnegative to control for the extreme case where the on-the-job training impact of rookies exceeds the available time of experienced personnel.[3]

Mentoring does not affect the steady state rookie fraction (eq. 13), as the flow of people through the experience chain remains the same. Mentoring, however, does lower average productivity:

Average Productivity = ((1 – Rookie Fraction) + (Rookie Productivity Fraction –
Fraction of Experienced Time Required for Training) * Rookie Fraction). (16)

The effect of mentoring on average productivity is proportional to the number of rookies that need to be trained. Thus, it can have a dramatic impact in situations where the rookie fraction is high, when the organization is growing rapidly. Figure

[3] A more robust formulation would gradually reduce the mentoring rookies receive as the workload of experienced employees grows, lengthening the assimilation time for rookies.

4 shows a simulation with the same parameters used in Figure 3b, but now assuming each rookie requires mentoring by the equivalent of 0.4 experienced people. Average steady state productivity drops from 0.65 to 0.45 FTEs per employee.[4]

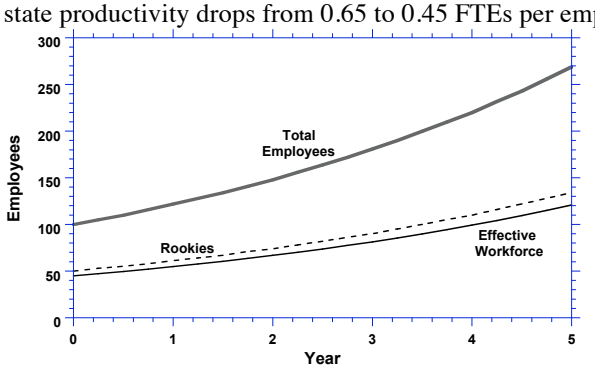


Figure 4. Effect of mentoring on workforce productivity.

Task processing and work pressure

In this section we present the structure for the arrival, accumulation and processing of customer orders and link it to the service capacity sector.

Tasks accumulate in a backlog until they are processed and delivered to the customer (Figure 5). The service backlog could be pending loan applications, tasks in a consulting project, the inbox of any administrative process, a physical queue of customers awaiting service at a bank or doctor's office, or the number of customers on hold at a call center. For now we assume an exogenous arrival rate.

$$\text{Service Backlog} = \text{INTEGRAL}(\text{Task Arrival Rate} - \text{Task Completion Rate}, \text{Service Backlog}_{t_0}). \quad (17)$$

Following Little's law, the average delivery delay (service time) is the ratio of the backlog to the completion rate:

$$\text{Delivery Delay} = \text{Service Backlog} / \text{Task Completion Rate}. \quad (18)$$

The completion rate is the lesser of (i) the potential completion rate based on the effective workforce (eq. 14) or (ii) the maximum completion rate, based on the number of tasks in the backlog and the minimum time needed to process each task.

$$\text{Task Completion Rate} = \text{MIN}(\text{Maximum Completion Rate}, \text{Potential Completion Rate}) \quad (19)$$

$$\text{Maximum Completion Rate} = \text{Service Backlog} / \text{Minimum Delivery Delay} \quad (20)$$

[4] An interactive version of this structure is available for experimentation at http://forio.com/resources/learning-curve-for-service-organizations/.

Potential Completion Rate =
Service Capacity * Standard Workweek / Standard Time per Task (21)

Service Capacity = Effective Workforce. (22)

The desired completion rate depends on the backlog and the organization's delivery time goal:

Desired Completion Rate = Service Backlog / Target Delivery Delay. (23)

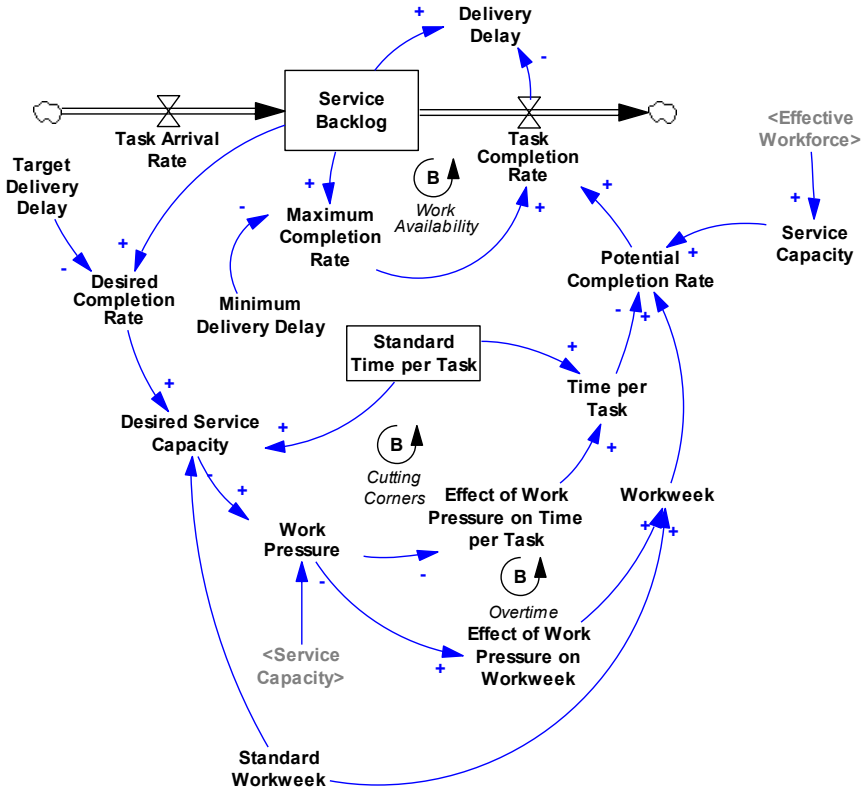


Figure 5. Feedbacks from Employee's responses to work pressure.

The organization must adjust service capacity (measured in person-hours of work effort available per week) to complete tasks at the desired rate. We define work pressure as the ratio of desired to actual service capacity. Desired capacity depends on the desired completion rate, the standard workweek and the standard time required to complete each task:

Work Pressure = Desired Service Capacity / Service Capacity (24)

Desired Service Capacity = Desired Completion Rate * Standard Time per Task / Standard Workweek. (25)

Work pressure greater than one indicates the service center is under stress as there are more tasks in the backlog than the center can process within the target delivery delay, given the number and productivity of employees, the standard workweek and the current standard for the time that should be allocated to each task. Work pressure less than one indicates excess capacity.

High work pressure should signal management to increase service capacity. However, service capacity responds with long lags: management must recognize the increase in work pressure, decide it is large and persistent enough to justify capacity expansion, authorize the new positions, then recruit, select, hire and train the new employees—and acquire the complementary capital stocks they need to become effective (office space, IT infrastructure, etc.). Until this occurs, employees are forced to handle high work pressure by either working harder (longer hours, fewer breaks) or by cutting corners (spending less time with each customer than needed to provide high quality service).

Employee responses to work pressure: working overtime and cutting corners

While management responds slowly to changes in work pressure, employees usually respond quickly: a bank teller sees the line of customers; a call-center representative knows when people are on hold; engineers know when their designs are late—and all know they must quickly boost throughput.

The first option for service providers facing high work pressure is to increase work intensity, that is, to work harder through overtime and by cutting the number and length of breaks. Thus, workweek is an increasing function of work pressure:

Workweek = Standard Workweek * Effect of Work Pressure on Workweek (26)

Effect of Work Pressure on Workweek = f(Work Pressure). (27)

Service providers can also respond to high work pressure by reducing the time per task. Speeding up might be as simple as reducing the time spent in pleasantries with the customer, but often involves "cutting corners", for example, failing to provide informative responses to customer queries, offer ancillary services, collect relevant information from the customer, check for errors, document the work or complete required reports. Time per task falls as work pressure increases:

Time per Task =
Standard Time per Task * Effect of Work Pressure on Time per Task (28)

Effect of Work Pressure on Time per Task = f(Work Pressure). (29)

The potential completion rate, eq. 21, now depends on service capacity modified by work intensity and actual time per task:

Potential Completion Rate = Service Capacity * Workweek / Time per Task. (21')

These responses create balancing feedbacks through which servers attempt to keep work pressure within certain limits. High work pressure boosts task completion through greater work intensity, reducing the backlog and work pressure (the Overtime loop in Figure 5). Similarly, high work pressure leads servers to cut corners, reducing time per task and speeding task completion, thus reducing the service backlog and work pressure (the Cutting Corners loop in Figure 5).

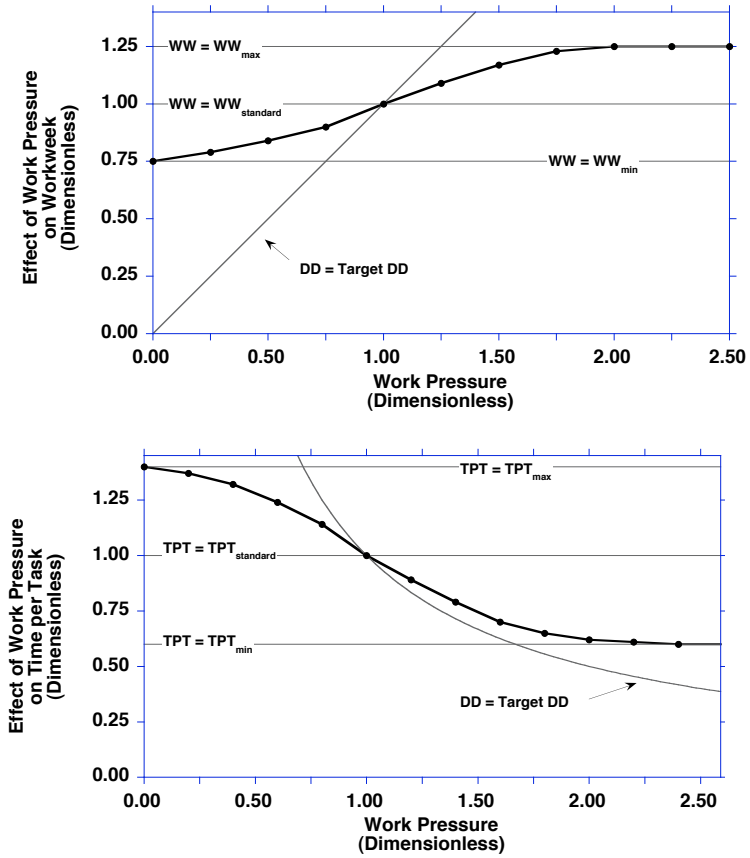


Figure 6. Employee's responses to work pressure

Figure 6 shows the effect of work pressure on workweek and time per task estimated from a detailed field study of a retail lending operation in a UK bank (see Dogan, 2007; Oliva, 2001; Sterman, 2000, §14.3, for details on the estimation process). In that bank, management required all tasks to be processed within one day. Employees used both overtime and corner cutting to meet this goal. Interestingly, the data show employees were twice as willing to cut corners as to work overtime. Note that both workweek and time per task saturate at extreme levels of work pressure: work hours cannot be increased beyond some level, and time per

task cannot be cut below some minimum level, even when work pressure is very high; similarly, the workweek does not fall to zero and time per task reaches some maximum even when work pressure is very low.

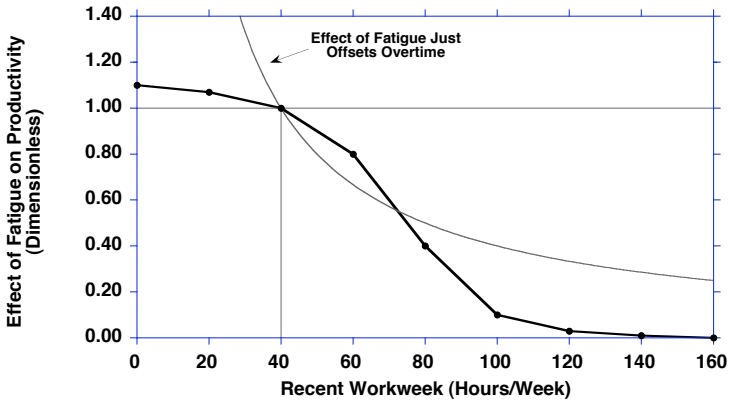


Figure 7. Effect of fatigue on productivity.

Side effects of overtime: fatigue and burnout

While higher work intensity boosts output in the short run, extended overtime causes fatigue that eventually undermines the benefit of longer hours (Homer, 1985; Thomas, 1993):

Service Capacity = Effective Workforce * Effect of Fatigue on Productivity. (22')

The effect of fatigue on productivity is a decreasing function that reduces service capacity when the recent workweek is greater than 40 hours/week, but increases only marginally above its normal operating point when the workweek falls below normal (Figure 7). Fatigue builds up and dissipates over time; we model fatigue as an exponentially weighted moving average of past work intensity. The longer the fatigue onset time the longer it takes for burnout to set in and for employees to recover when work intensity falls.

Effect of Fatigue on Productivity = f(Recent Workweek) (30)

Recent Workweek = SMOOTH(Workweek, Fatigue Onset Time) (31)

where Output = SMOOTH(Input, Averaging Time) denotes first-order exponential smoothing of the input with a mean delay of the Averaging Time (see Sterman 2000, ch. 11 for details).

Extended periods of high work intensity also increase employee turnover. The effect of fatigue on turnover is an increasing function of burnout, and affects both types of employees:

Rookie Quit Rate = Rookie Employees * Rookie Quit Fraction * Effect of Fatigue on Turnover (4')

Experienced Quit Rate = Experienced Employees * Experienced Quit Fraction * Effect of Fatigue on Turnover (5')

Effect of Fatigue on Turnover = f(Long Term Workweek). (32)

Long Term Workweek = SMOOTH(Workweek, Burnout Onset Time). (33)

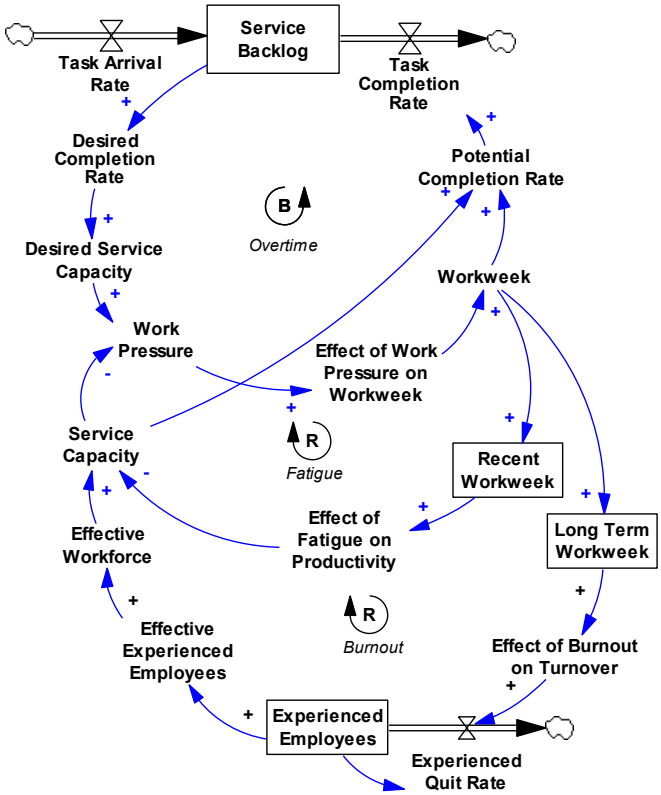


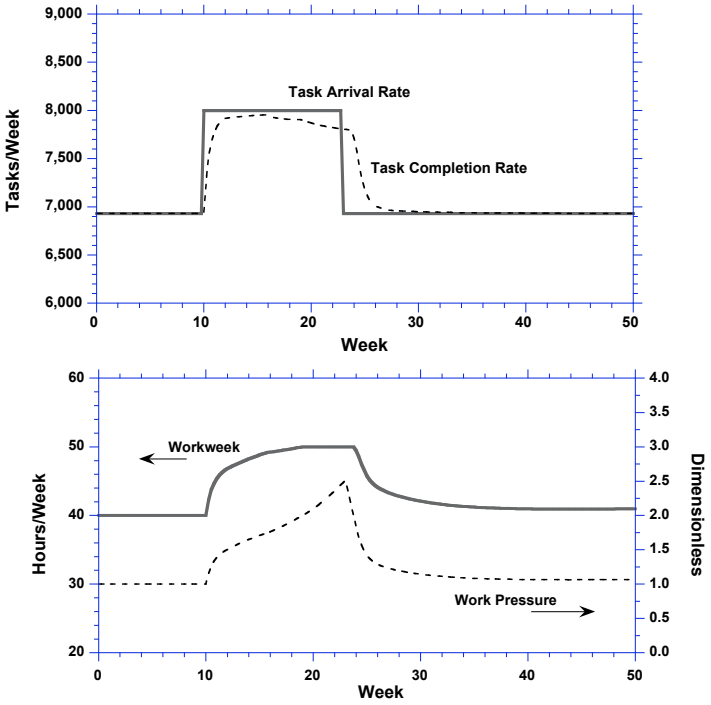
Figure 8. Consequences of sustained work intensity.

Like the effect of fatigue on productivity, extended overtime increases attrition with a delay, but with a longer time constant: long workweeks quickly reduce productivity, but people will tolerate high overtime much longer before quitting.

These two 'side effects' of high work intensity create a pair of reinforcing feedbacks that can trap the organization in substandard performance. Fatigue and burnout reduce service capacity (directly and indirectly, as attrition both lowers headcount and increases the rookie fraction), which—*ceteris paribus*—lowers task completion, pushing the service backlog up, further increasing work pressure and

forcing service providers to work even harder (the Fatigue and Burnout loops in Figure 8).

To illustrate, Figure 9 shows the impact of a 15% increase in task arrivals for one quarter, starting from equilibrium and assuming workweek is the only adjustment process (holding time per task constant). Hiring is set to replace total quits, so the workforce remains constant throughout. When arrivals increase, the backlog and work pressure grow, and employees immediately increase their workweek. Task completion rises, though not enough to match the arrival rate. Backlog continues to accumulate, and work pressure grows further. By week 15 the effects of fatigue overcome the benefits of longer hours and the completion rate begins to drop. By week 20 employees reach the maximum possible workweek (see Figure 6). In week 23 the arrival rate drops back to its original level. Work completion



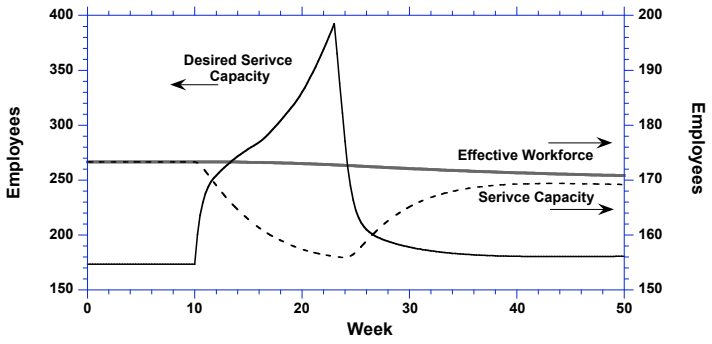


Figure 9. Work intensity response.

gradually falls as the backlog drops, reaching its desired level in approximately 5 weeks. Note, however, that the system settles into a new equilibrium. Burnout from extended overtime increased turnover, shifting the employee mix to include more rookies. With the base case parameters the steady-state rookie fraction rises from 16.6% to 18.2%, causing a 1.4% drop in average productivity. As a result, work pressure does not return to normal: the same number of less productive employees are forced to maintain a slightly longer workweek. A *temporary* surge in work volume caused a *permanent* drop in productivity.

Side effects of cutting corners: lower quality and standard erosion

While cutting corners immediately increases output, it does so at the cost of the quality of the customer's experience and a higher likelihood of errors. We begin with the impact of corner cutting on service operations; below we consider how corner-cutting feeds back to affect the firm's competitiveness and customer base.

Effects of lower quality

A common way to cut the time spent on each task is to skip steps and cut quality assurance. The obvious unintended impact is a higher error rate, leading to customer dissatisfaction and costly rework. Errors are typically not detected immediately: A waiter in a hurry may take the customer's lunch order without reading it back for confirmation, but the error is not discovered until the customer receives the tuna surprise instead of the tofu burger. Credit card billing errors are typically discovered only after customers examine their monthly statements. Errors therefore accumulate in a stock of undiscovered rework until they are discovered (Lyneis and Ford, 2007; Sterman, 2000, ch. 2).

$$\text{Undiscovered Errors} = \text{INTEGRAL}(\text{Error Generation Rate} - \text{Error Discovery Rate}, \text{Undiscovered Errors}_{t_0}). \quad (34)$$

The error discovery rate is assumed to be a first-order process with a constant average error discovery time:

Error Discovery Rate = Undiscovered Errors / Time to Discover Errors. (35)

Error generation depends on the total completion rate and the probability that each task contains an error:

Error Generation Rate = Task Completion Rate * Probability of Error Generation. (36)

We assume that the probability of errors depends on three factors: corner cutting (time per task), fatigue, and average employee experience. Cutting the time spent on each task increases the probability of error as employees hurry, skip steps, and fail to check their work. Fatigue increases the chance of error and cuts the chance of detecting and correcting it at the time. Inexperienced personnel make more errors. For simplicity we assume these sources of error are independent. The probability a task is done incorrectly is then the complement of the probability that no error was introduced by any of these factors:

$$\text{Probability of Error Generation} = 1 - \prod_{i \in \{F\}} \text{Probability of Error Free}_i \quad (37)$$

$$\text{Probability Error Free}_i = f(\text{F}_i) \quad (38)$$

where $F \in$ {Time per Task, Recent Workweek, Average Productivity}.

When errors are discovered they are added to the service backlog to await reprocessing. Thus, equation 17 is modified to

Service Backlog =
INTEGRAL(Task Arrival Rate + Error Discovery Rate – Task Completion Rate,
Service Backlog$_{t_0}$). (17')

Corner cutting also influences employee attrition. Employees will endure more pressure and develop greater loyalty to the organization if they perceive that they deliver a high-quality service (Schneider, 1991; Schneider et al., 1980). Alternatively, if employees perceive low levels of service quality they are more likely to leave the organization. The effect of quality on turnover modifies the fractional attrition rate for all employees:

Rookie Quit Rate = Rookie Employees * Rookie Quit Fraction *
Effect of Fatigue on Turnover * Effect of Quality on Turnover (4'')

Experienced Quit Rate = Experienced Employees * Experienced Quit Fraction *
Effect of Fatigue on Turnover * Effect of Quality on Turnover. (5'')

The effect of quality on turnover is modeled as an increasing function of the quality perceived by employees. Employees are assumed to adjust their perception of service quality after a short delay. We model this perception process as a first order exponential average of the actual quality delivered:

$$\text{Effect of Quality on Turnover} = f(\text{Perceived Quality}_E) \quad (39)$$

$$\text{Perceived Quality}_E = \text{SMOOTH}(\text{Delivered Quality, Time to Perceive Quality}_E). \quad (40)$$

Service quality is, by definition, determined by the customer's subjective experience with the service organization. We model quality as a function of the performance gap—the difference between the time allocated per task and the customer's expectation of what that time should be. Delivered quality is one when the performance gap is zero, that is, when time per task matches customers' expectations (Figure 10). The existence of a "tolerance zone" for service quality (Strandvik, 1994; Zeithaml et al., 1993) suggests that the function is relatively flat when the performance gap is small, but grows progressively steeper with the gap.

$$\text{Delivered Quality} = f(\text{Time per Task} - \text{Customer Expected Time per Task}) / \text{Customer Expected Time per Task}). \quad (41)$$

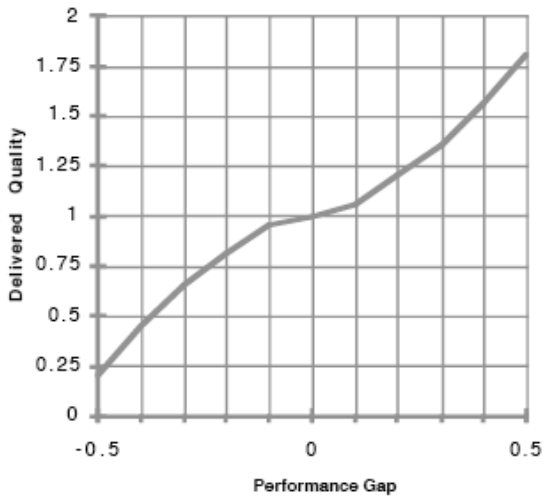


Figure 10. Effect of preformance gap on delivered quality.

The introduction of rework creates another performance trap. Corner cutting eases high work pressure, but also causes quality to drop while increasing errors. When the errors are discovered they must be reworked, further increasing work pressure and pushing employees to cut corners still more (the Rework loop in Figure 11). Low quality also boosts attrition, reducing average productivity, creating another positive feedback (the Disappointment loop in Figure 11).

The unintended consequences of corner cutting are similar to those of increased work intensity: the effect of quality on turnover is structurally identical to the effect of burnout on turnover and the effect of errors similar to the productivity losses from fatigue. The strength and time constants for these effects differ from the

workweek impacts analyzed in Figure 9, but the resulting behavior is qualitatively similar. After a temporary increase in task arrivals, the system reaches equilibrium with more inexperienced employees, causing sustained work pressure. The higher rookie fraction is the result of the increased turnover caused by lower quality. With fewer experienced people, work pressure remains above normal, leading to more errors and lower service quality.

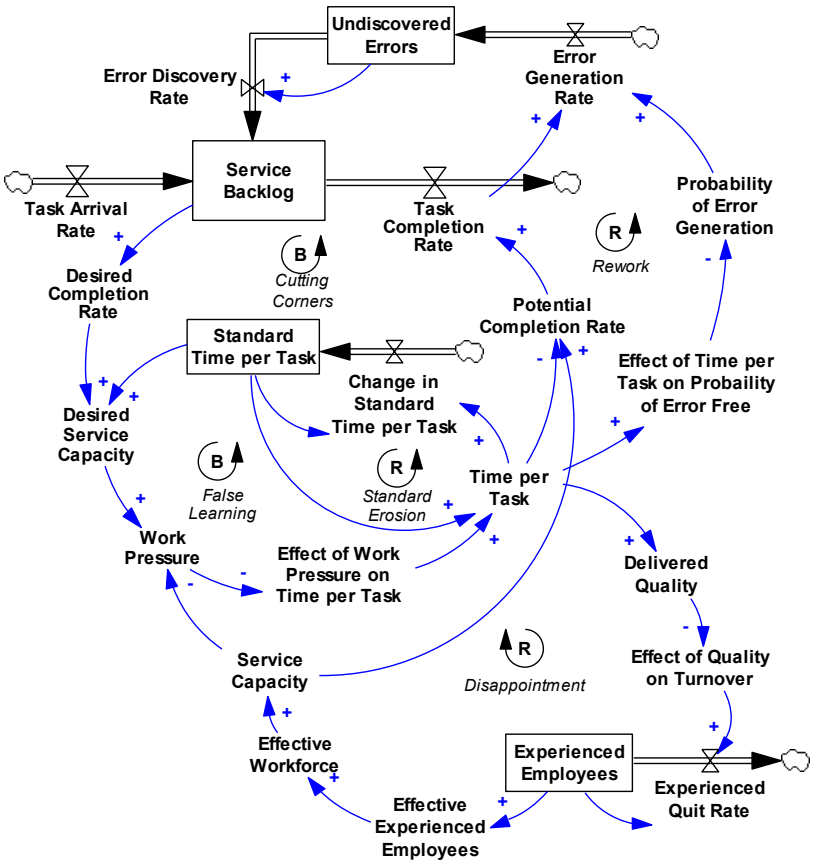


Figure 11. Consequences of sustained corner cutting.

Erosion of service standards

Services are intangible and quality is difficult to measure. In the absence of compelling external feedback on service quality, an organization's internal standards for service quality tend to drift with past performance. The expectation formation literature suggests that performance standards are adjusted based on an an-

choring and adjustment heuristic (Lant, 1992; Lewin et al., 1944). We model the adjustment process for the standard time per task, the time employees would allocate to each task in the absence of work pressure, as an asymmetric process. Asymmetric adjustments have been used in the organizational and psychological literature to represent biased formation of expectations and goals, and are normally formulated with different time constants governing the adjustment process depending on whether the aspiration level is above or below actual performance:

$$\text{Standard Time per Task} = \text{INTEGRAL}((\text{Time per Task} - \text{Standard Time per Task}) / \text{Time to Adjust Standard}, \text{Standard Time per Task}_{t_0}) \quad (42)$$

$$\text{Time to Adjust Standard} = \text{IF}(\text{Time per Task} < \text{Standard Time per Task}, \text{Time to Adjust Down}, \text{Time to Adjust Up}). \quad (43)$$

Oliva and Sterman's (2001) financial services field study showed that the organization's standard time per task adjusted downward much faster than it adjusted upward. Management interpreted any reduction in time per task as cost-saving productivity improvement rather than as a sign of poor quality. Upward adjustments of standard time per task, in contrast, were resisted as they imply a reduction in productivity. Indeed, Oliva and Sterman found that the best estimate of the downward adjustment time for quality standards was 19 weeks, while the upward adjustment time was essentially infinite—temporary corner cutting was quickly embedded in organizational norms for standard time per task, while increases in time per task did not result in upward revision of quality norms.

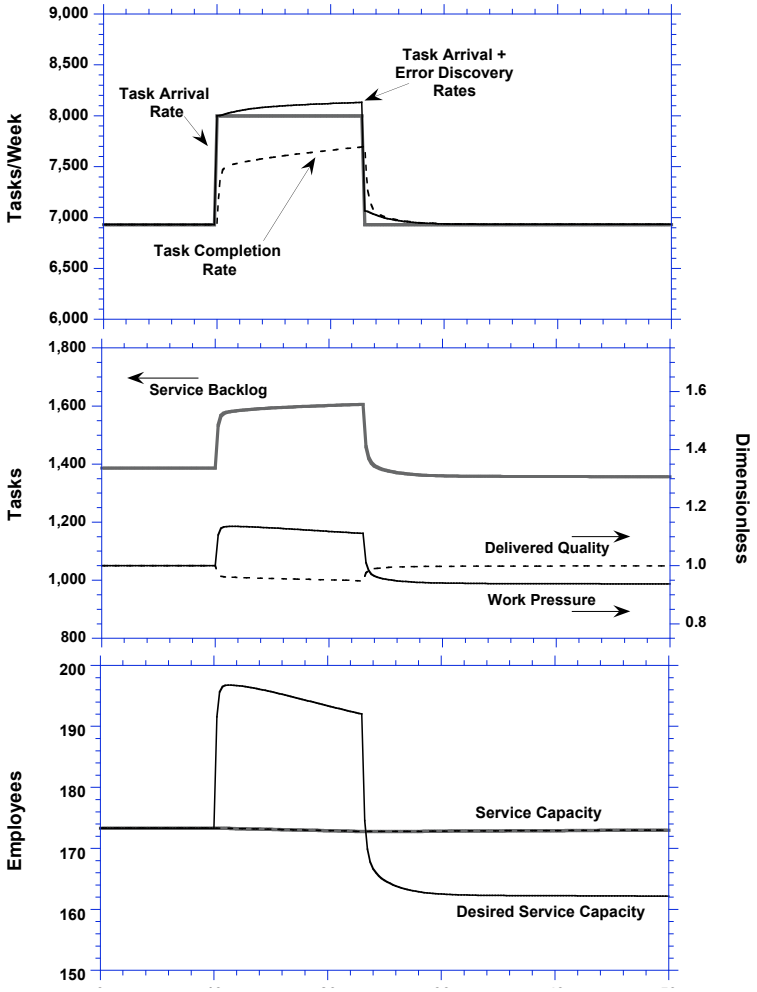
Incorporating dynamic quality norms creates a new reinforcing loop that can trap the firm in substandard performance. As high work pressure causes temporary corner cutting, standard time per task begins to fall (eq. 42). If work pressure remains high, however, employees seeking to clear the backlog faster will cut time per task below the new, lower standard, causing still more erosion of the quality norm (the Standard Erosion loop in Figure 11).

Furthermore, since management uses the standard time per task to estimate required service capacity (eq. 25), reductions in the standard, *ceteris paribus,* reduce required capacity, thus easing work pressure. That is, management interprets erosion in the time needed per task as permanent productivity improvement due to learning (the False Learning loop in Figure 11). Goal erosion provides another negative feedback through which high work pressure can be eliminated.

Figure 12 shows the impact of corner cutting and standard erosion, holding the workweek constant (thus eliminating the impact of overtime on task completion, errors, productivity, and turnover). The system begins in equilibrium with enough capacity to deliver high quality work with no errors. In week 10 task arrivals rise by 15% for one quarter, then fall back to the original value.

The surge in arrivals causes the backlog to grow, increasing work pressure. Employees respond by reducing time per task, but the completion rate remains below arrivals, so the backlog continues to grow. The unanticipated side effects of corner cutting soon appear: less time per task increases errors; after a delay, these are discovered, increasing the backlog still further. Low quality increases employee turnover, reducing effective capacity as rookies replace experienced staff.

Lower effective capacity forces work pressure up still more. Finally, the standard for time per task gradually falls as workers and management become habituated to spending less time with each customer, cut quality checks, reduce effort to understand the customers' needs and cross-sell additional services, and fail to document their work. Eroding quality standards eases work pressure even though the service backlog continues to grow (the False Learning loop in Figure 11).



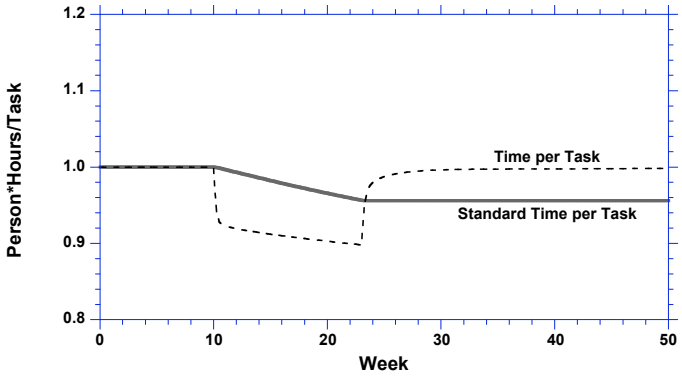


Figure 12. Consequences of sustained corner cutting

When task arrivals fall back to the original value, the backlog and work pressure quickly fall, and time per task increases. In the new equilibrium work pressure is less than one, indicating excess capacity. Standard time per task fell during the period of high work pressure, but does not rise when work pressure is low. Since headcount remains constant throughout the simulation, lower standards (higher perceived employee productivity) mean service capacity eventually exceeds demand. With low work pressure, employees can restore service quality close to its original level. However, in reality, management would not long tolerate such excess capacity.

Management response to work pressure: adjusting service capacity

In reality, management is likely to respond to imbalances between desired and actual service capacity by altering the workforce. If the workforce could adjust quickly and fully in response to changes in required capacity, overtime and corner cutting, with their unintended harmful consequences, would be minimized. However, expanding the workforce is expensive and time consuming, and it is costly and disruptive to reduce headcount. Managers of service operations often face severe budget constraints and pressure to meet financial targets. Capacity expansion is, therefore, often the response of last resort.

To capture capacity adjustment endogenously we now replace the constant-headcount hiring policy (eq. 11) with a more realistic decision rule that adjusts the workforce in response to the gap between desired and actual service capacity. First, hiring takes time—time to create and advertise vacancies, review applicants, interview candidates, and fill positions. The difference between the rates at which new positions are authorized and filled accumulate in a stock of unfilled vacancies. We assume the average time to hire is constant (in reality it varies with labor market conditions, rising when labor markets are tight and falling when unemployment is high):

Rookie Hiring Rate = Employee Vacancies / Time to Hire (11')

Employee Vacancies = INTEGRAL(Labor Order Rate – Rookie Hiring Rate,
Desired Vacancies). (44)

Orders for labor – the rate at which vacancies are created – are normally determined by the desired hiring rate corrected for any discrepancies between the desired and actual number of vacancies. If, however, there were an extreme surplus of workers labor orders could become negative, forcing existing job openings to be canceled. In such a situation, vacancy cancellation is constrained to be no faster the rate determined by the average time required to cancel open vacancies:

Labor Order Rate =
MAX(Desired Hiring Rate + (Desired Vacancies – Vacancies) / Time to Adjust Workforce, – Employee Vacancies / Time to Cancel Vacancies). (45)

The number of vacancies needed to hire at the desired rate is, by Little's Law, proportional to the desired hiring rate and average delay in filling vacancies:

Desired Vacancies = Desired Hiring Rate * Time to Hire. (46)

The organization seeks to replace those employees who have quit and correct any discrepancy between desired and existing labor. The responsiveness of the policy is given by the time to adjust the workforce:

Desired Hiring Rate = Total Quit Rate + Workforce Adjustment Rate (47)

Workforce Adjustment Rate =
(Desired Workforce – Total Employees) / Time to Adjust Workforce. (48)

The desired workforce is determined by desired service capacity and management's belief about average productivity. However, because labor is costly and slow to change, management does not act instantaneously on labor requirements. Instead, the desired workforce adjusts with a lag to the level indicated by desired service capacity and perceived employee productivity. The lag, modeled here by first-order exponential smoothing, ensures that capacity and hiring do not overreact to temporary variations in service demand:

Desired Workforce =
SMOOTH(Desired Service Capacity / Perceived Employee Effectiveness,
Time to Adjust Desired Workforce). (49)

Furthermore, employee effectiveness is not perceived instantaneously, since it takes time to measure, report and assess changes in productivity. We model that process with exponential smoothing of actual employee effectiveness:

Perceived Employee Effectiveness = SMOOTH(Service Capacity / Total Employees,
Time to Perceive Productivity). (50)

Note that perceived employee effectiveness is an aggregate measure based on the data management actually has available: productivity is the ratio of service capacity (which is the task completion rate converted to labor requirements using the

standard work week and standard time per task) to total staff. Consequently, managers' beliefs about employee productivity adjust gradually to variations in productivity caused by changes in the rookie fraction, fatigue, and erosion of the standard time per task.

Management's decision rule for hiring acts to eliminate discrepancies between desired and actual service capacity, creating another negative feedback loop through which the organization can regulate work pressure.

Figure 13 shows the impact of all three ways workers and managers respond to work pressure: overtime, corner-cutting, and hiring. From an initial equilibrium there is a sudden, unanticipated, permanent 10% increase in task arrivals. Figure 13 shows the contribution to task completions from each response, along with the change in throughput resulting from service standard erosion. The combined responses are capable of immediately boosting task completions to match the arrival rate. However, the timing and strength of the responses differs substantially.

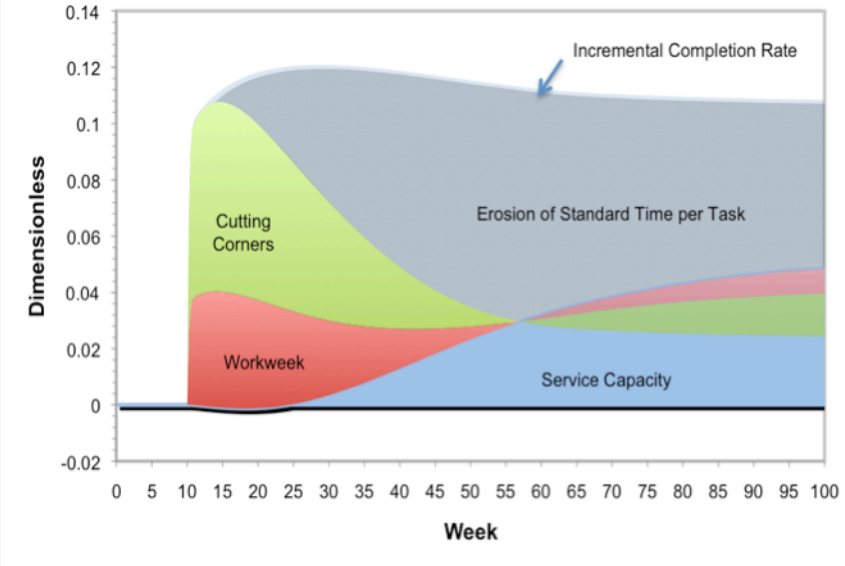


Figure 13. Response to a 10% increase in demand.

First, as estimated by Oliva and Sterman (2001) and shown in Figure 6, employees are twice as willing to cut corners as to work overtime. The reduction in time per task, along with fatigue and experience reduction, cause errors to grow; as these are discovered the task arrival rate rises beyond the 10% exogenous shock (eventually peaking 11.4% above the initial rate). At the same time, internal quality standards (standard time per task) begin to erode. Longer hours and lower quality cause higher turnover, lowering effective service capacity. Lower standards for time per task ease work pressure somewhat, but it remains well above normal.

The service standard continues to erode as employees respond to continued work pressure by still more corner cutting. Meanwhile, management responds to

the high workload by increasing the desired workforce, but the lags in recognizing and responding to the need, and in filling vacancies, mean service capacity begins to increase only after week 25. Service capacity reaches the desired level by week 58, then overshoots. Service capacity overshoots because of the delays in perceiving the adequacy of service capacity and in filling vacancies. Further, even as hiring slows, the many rookies hired in response to the demand surge continue to gain experience, raising effective capacity. Excessive capacity causes work pressure to fall below one. Employees then spend more time with each customer than the (now lower) service standard indicates, and reduce their workweek (taking longer breaks, using more work time for personal business, etc.). The inertia of the hiring process causes service capacity to peak two years after task arrivals increase, and the delays in the learning curve and in changing perceptions of productivity mean it takes almost five years for the system to return to equilibrium.[5]

Most important, the new equilibrium reached by the system is very different from the original equilibrium. While task arrivals rise by 10%, capacity does not expand by 10%. Rather, most of the growth in throughput results from a permanent reduction in the organization's internal quality standard. With the base case parameters, capacity expands in equilibrium by only 2.1%, with permanent standard erosion providing the rest of the "capacity" needed to meet the increase in demand. Note also that the drop in the equilibrium time per task causes a rise in errors and rework contributing an additional 1.3 percentage point increase in task completion compared to the original level, and that the lower quality increases employee turnover, causing the rookie fraction to increase by one percentage point, to 17.7%, lowering productivity and increasing costs.

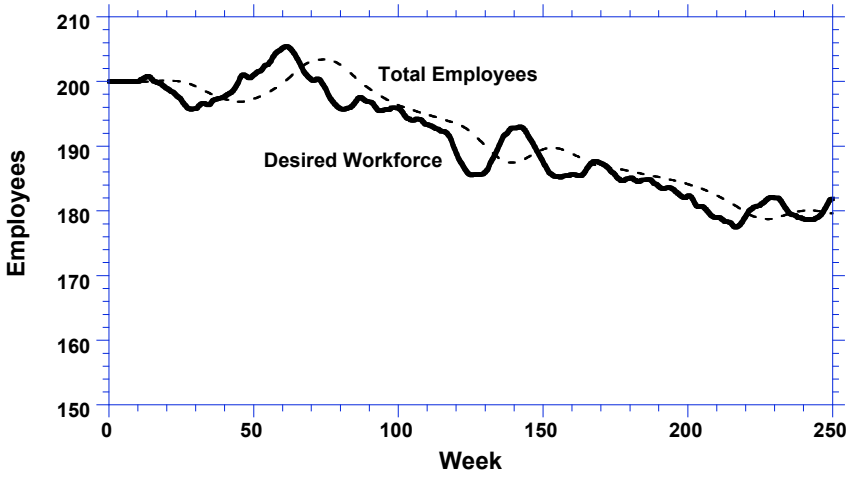
To assess system response under more natural conditions than a single increase in demand, we subject the model to random variations in the task arrival rate. We assume arrivals are determined by a pink noise process with a standard deviation of 5% and first-order autocorrelation time constant of four weeks (Figure 14). Most people intuitively believe that, since orders are stationary, the firm's initial resources should, on average, be sufficient to maintain the service standard and desired delivery delay. Instead, the system exhibits persistent erosion in the service standard (in this case, an average of 2.1%/year). The asymmetric adjustment of the standard time per task is responsible. When task arrivals exceed the mean, work pressure rises, time per task falls, and the standard drops a bit. However, when task arrivals are lower than the mean and work pressure falls below one, time per task rises, but the standard does not adjust upward. Management responds to the small, but cumulative, decrease in person-hours per task by gradually raising their estimate of workforce productivity, leading them to reduce desired service capacity accordingly. As service capacity falls, work pressure rises, which

[5] The adjustment to equilibrium is also slowed by the assumed low attrition rate of 20%/year and assumption that the firm does not lay off excess staff. Many service operations, particularly low-wage settings such as retail, entry-level financial services, fast food, and call centers experience far higher turnover. The model can also be easily expanded to allow for layoffs (Sterman 2000, ch. 19).

in turn leads to further corner cutting and standard erosion, thus locking the system into a vicious cycle.

Market feedback

Until now we have assumed task arrivals are exogenous, corresponding to a captive customer base. Such scenarios are approximated in many settings, for example, health care, financial services, and hardware help desks, where customers must seek medical care from the doctors in their existing health plan, file claims with their existing insurance policy, and seek service from the help desk of the firm from which they bought their new laptop. However, even in such captive situations, customers usually have the option of switching to other providers over the long term. We now expand the boundary of the model to incorporate the main market feedbacks that drive the customer base and task arrivals.



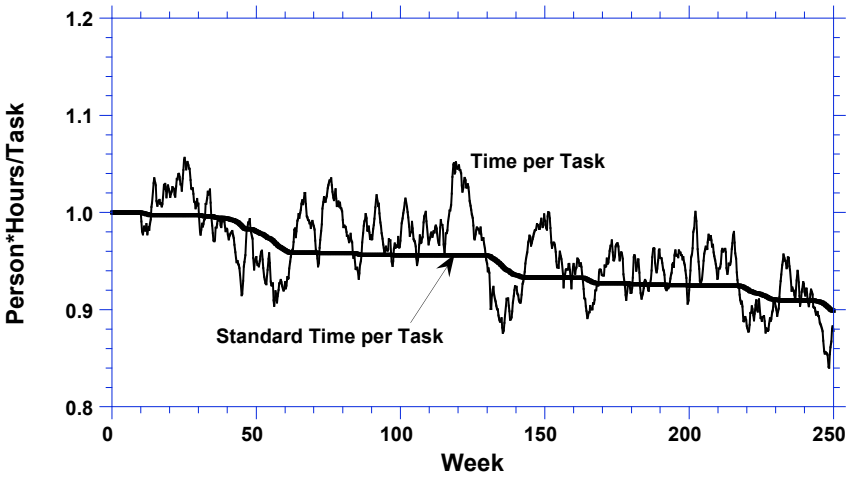


Figure 14. Response to stationary random orders with 5% standard deviation.

We assume tasks arrive at a rate proportional to the customer base:

$$\text{Task Arrival Rate} = \text{Customer Base} * \text{Task Requests per Customer}. \tag{51}$$

For simplicity we assume task requests per customer are exogenous. The customer base is formulated to increase at a rate that depends on perceived service attractiveness. When service attractiveness is greater (less) than one, the customer base will gradually rise (fall) as the firm wins or loses customers to competitors:

$$\begin{aligned}\text{Customer Base} = \text{INTEGRAL}(&\text{Base Customer Growth Rate} + \\ &(\text{Customer Base} * \text{Service Attractiveness} - \text{Customer Base}) / \\ &\text{Time to Adjust Customer Base}, \text{Customer Base}_{t_0})\end{aligned} \tag{52}$$

where the Base Customer Growth Rate is an exogenous fractional increase in customers, to allow for growth in the underlying market. Service attractiveness responds with a delay to the product of four attributes of the service encounter: errors, balking, the service delivery time, and delivered service quality. The delay represents the time required for customer beliefs about service quality to change, and for a new level of service attractiveness to persist long enough to induce customers to switch providers:

$$\begin{aligned}\text{Service Attractiveness} = \text{SMOOTH}(&\prod_{j\in\{A\}} \text{Effect of Attribute}_j \text{ on Attractiveness}, \\ &\text{Time to React to Service Attractiveness})\end{aligned} \tag{53}$$

$$\begin{aligned}\text{Effect of Attribute}_j \text{ on Attractiveness} = \\ (\text{Actual Performance Attribute}_i / \text{Standard Attribute}_i)^{\text{Sensitivity of Attractiveness from Attribute}_i}\end{aligned} \tag{54}$$

where Attribute $A_i \in$ {Error Discovery Rate, Balking Rate, Delivery Delay, Delivered Quality}.

Finally, the balking rate—the rate at which customers abandon the service backlog because of excessive waiting time—depends on an increasing function of the delivery delay relative to the customers' standard for wait time:

Balking Rate = Service Backlog * Normal Balking Rate *
Effect of Delivery Delay on Balking (55)

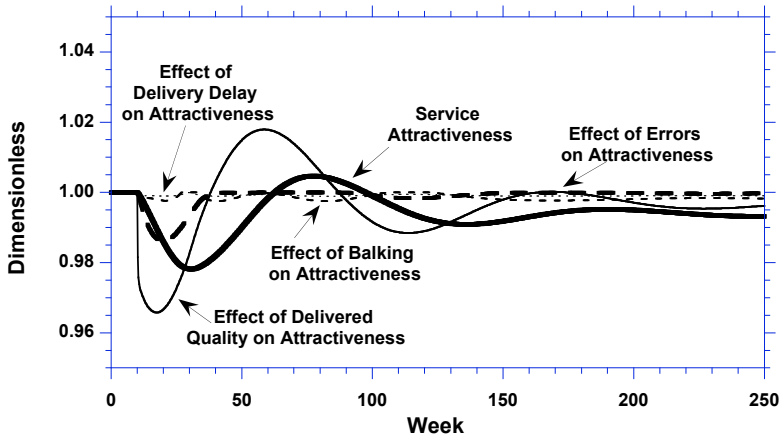
Effect of Delivery Delay on Balking =
f(Delivery Delay / Customer Standard for Delivery Delay). (56)

The service backlog is now:

Service Backlog =
INTEGRAL(Task Arrival Rate + Error Discovery Rate –
Task Completion Rate – Balking Rate, Service Backlog_{t_0}). (17")

For simplicity, we assume that customers who balk do not return to the queue at a later time. However, the higher the rate of balking, the lower are customer perceptions of service quality (eq. 53), which feed back to the customer base (eq. 52).

Figure 15 shows the service center's response to a 10% increase in tasks requested per customer. As before, the surge in workload leads to overtime and corner cutting. These responses, along with eventual hiring, together allow the service center to process the higher load in the normal delivery time, so the impact of wait time on balking and customer perceptions of service attractiveness is minimal.



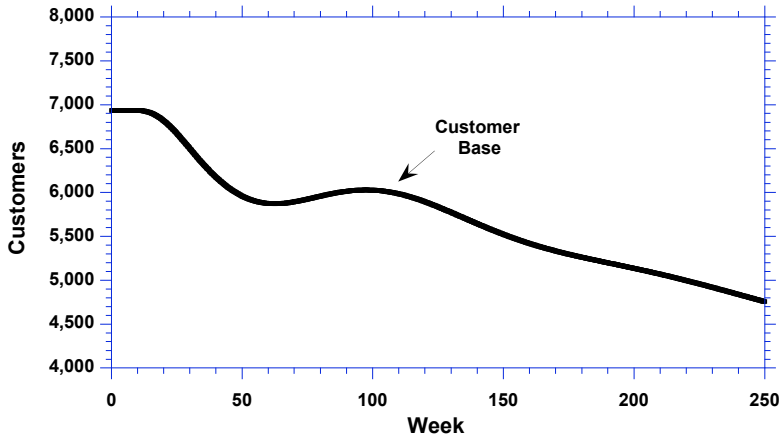


Figure 15. Response to a 10% increase in demand with market feedback.

However, overtime and corner cutting increase errors and lower the quality of service the customers experience. It takes time for customers to perceive the drop in service attractiveness, but as they do, the customer base begins to erode.

The drop in customer base and slight increase in service capacity eventually bring capacity in line with service demand, and attractiveness returns to normal (week 35). With a delay, customers react to the improvement in quality, and the customer base stabilizes, which then starts to rise as excess service capacity temporarily improves quality. However, the reduction in time per task during the period of high work pressure caused throughput per worker to rise. Observing this increase in task completions per person, management raises their estimate of worker productivity (as in Figures 12 and 14), leading desired service capacity to fall more than the drop in work volume arising from the erosion of the customer base. The resulting drop in capacity then raises work pressure, leading to additional corner cutting and lower quality, further eroding the customer base. The organization is captured in a vicious cycle: high work pressure erodes service standards and lowers capacity, ensuring that work pressure remains high and standards continue to erode.

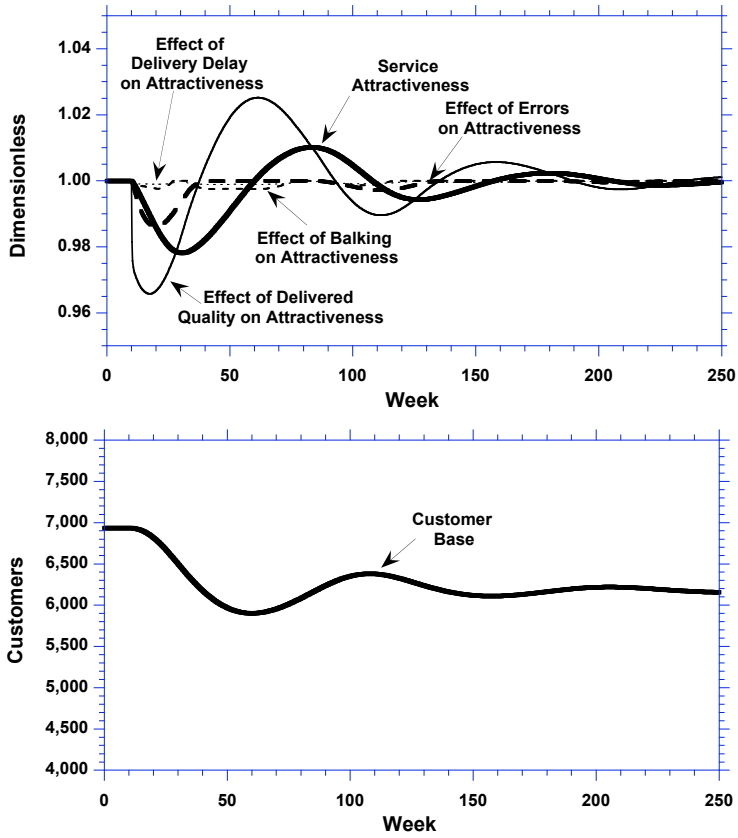


Figure 16. Response to a 10% increase in demand with market feedback and the possibility to improve Standard Time per Order

As discussed above, these results could be in part explained by the asymmetric adjustment of the service standard documented in Oliva and Sterman (2001): standards can fall, but not rise. To test the sensitivity of the results to this assumption we ran the same test as before, but allowing standard time per task to increase when time per task rises above the standard. We set the time constant for upward adjustment equal to 150% of the value for downward adjustment—an optimistic estimate of managements' ability to recognize the benefits of the increased level of service and to build them into organizational practices. As in the previous scenario, the surge in demand triggers overtime, corner cutting and standard erosion, leading to a drop in the customer base (Figure 16). Unlike the previous case, however, during the period of excess service capacity, the service standard rebounds. The system reaches an equilibrium in which service capacity matches demand and employees deliver the standard time per task working the standard week. Note that while the standard time per task returns to its original value, it

does so only after the firm permanently loses nearly 10% of its customer base. Upward adjustment of the quality standard allows the firm to halt the death spiral, but, rebuilding the customer base would require the firm to increase service quality and standards above the original levels; the interactions of the routines for assessing worker productivity and hiring do not create such conditions.

Financial pressure

Up to now the organization has been free to hire as many people as it deems necessary to meet demand. We now expand the model boundary to include financial constraints on hiring. To do so, we revise the workforce adjustment to respond to the authorized workforce, defined to be the lesser of the workforce needed to meet demand or what the organization can afford given its budget:

Workforce Adjustment Rate =
    (Authorized Workforce – Total Employees) / Time to Adjust Workforce (48')

Authorized Workforce = MIN(Desired Workforce, Affordable Workforce) *
    Margin for Reserve Capacity. (57)

The margin for reserve capacity represents a nominal fractional level of excess capacity built into budgets and staffing. A margin of 1+m indicates that the staff target for the service organization is m% higher than the level just sufficient to meet demand, and that the budget to cover this reserve capacity is also made available. The desired workforce continues to be driven by throughput requirements described above. The affordable workforce is determined by the operating budget and fully-loaded cost per employee. Again, because labor is costly and slow to change, management does not react instantaneously to budget changes. Like the desired workforce, the affordable workforce adjusts to the level indicated by the budget and cost per employee with a lag:

Affordable Workforce = SMOOTH(Budget / Salary per Employee,
    Time to Adjust Affordable Workforce). (58)

The operating budget is assumed to have a fixed component and a component dependent on generated revenues:

Budget =
    MAX(0, Base Budget + Revenue * Fraction of Revenue for Operations). (59)

The budget formulation allows a wide range of service organizations to be modeled. A single customer support call center in a firm may generate no revenue; such centers are typically managed as cost centers and must live within a given base budget each year. At the other extreme, when the model represents an entire firm, the budget must then come (nearly) entirely from revenue.

For simplicity we model revenue as proportional to the customer base and the average revenue generated per customer per month, independent of the number of service requests each customer generates, an approximation of many settings in

which customers pay a certain monthly fee, such as insurance premium payments or account maintenance fees in financial services:

Revenue = Customer Base * Revenue per Customer. (60)

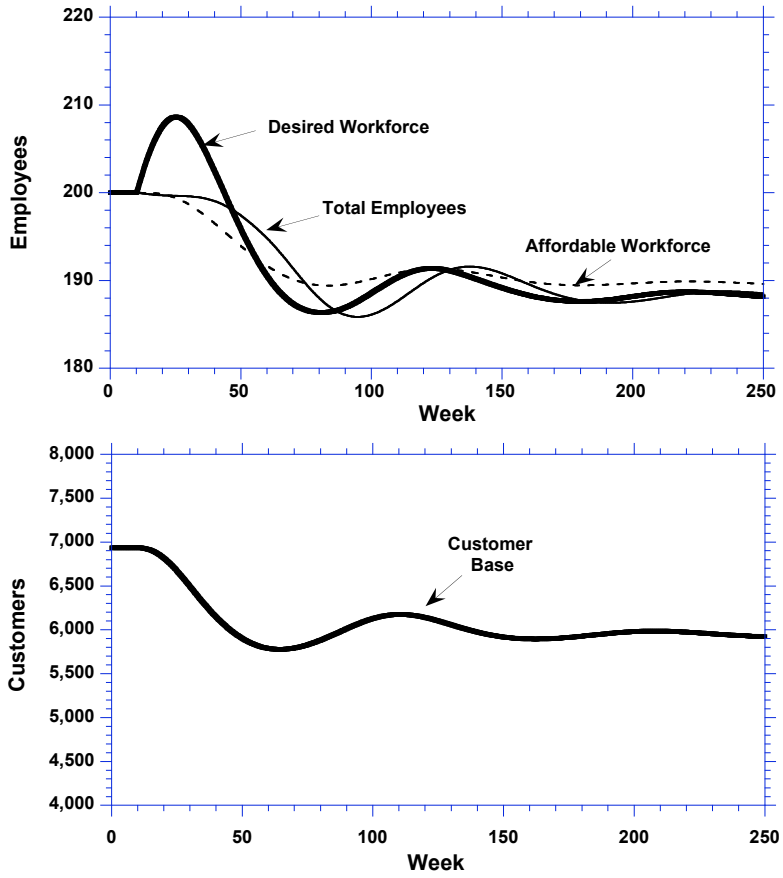


Figure 17. Response to a 10% increase in demand with market feedback, the possibility to improve Standard Time per Order, and financial constraints on hiring

For the base case, we start the system in equilibrium with the budget set to cover the cost of the work force exactly, i.e., no margin for reserve capacity. Figure 17 shows the evolution of the desired, affordable and actual workforce for the same scenario as Figure 16, but with the budgeting process active. The financial constraint means the workforce does not rise after the surge in demand. As before, the increase in work pressure causes overtime and corner cutting that reduce service attractiveness, causing a gradual drop in the customer base. Now, however, the drop in revenue caused by the loss of customers forces the workforce down as

the budget falls. Capacity remains inadequate, forcing employees to cut corners further. The customer base drops still more. Like the simulation in figure 16, the system reaches equilibrium with attractiveness returning to normal. However, because the budget constrained hiring during the transient, the customer base drops 14.6% instead of 8.9%. The death spiral in this simulation halts only because the base budget of the service center creates a floor for the affordable workforce. Results are worse if the organization must rely on revenue for its budget to a greater degree.

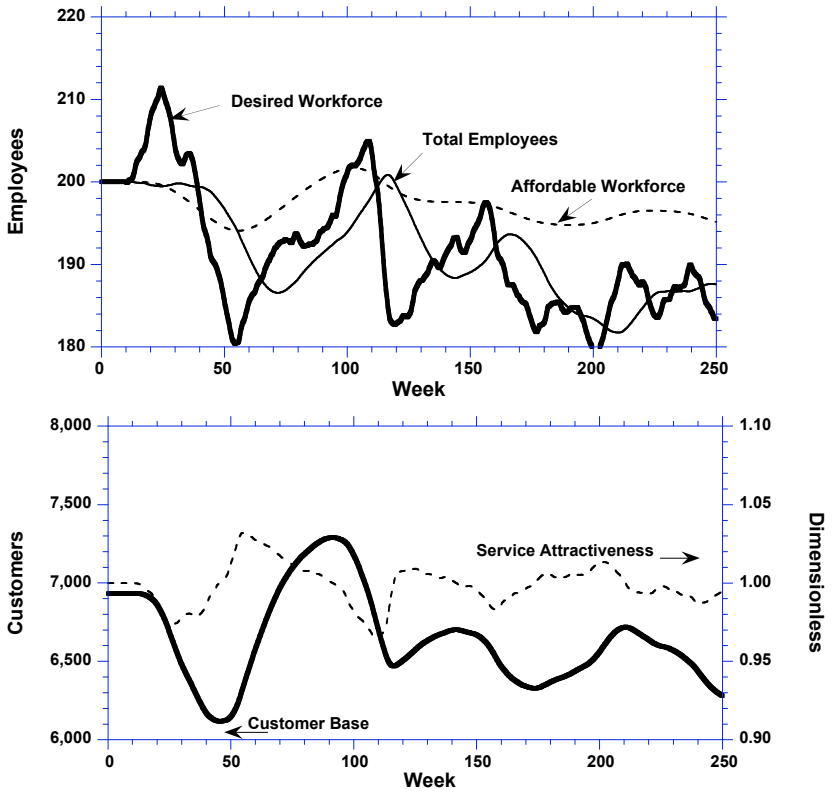


Figure 18. Response to stationary random orders with 10% standard deviation – Full model (base case)

To assess the full system response under more realistic conditions, we subject it to random variations in the task arrival rate. We assume arrivals are determined by a pink noise process with a standard deviation of 10% and first-order autocorrelation time constant of four weeks. The system begins in equilibrium with a time constant to adjust the standard time per task upward equal to 150% of the downward adjustment time. While the system is capable of maintaining service attractiveness very close to its normal operating point, it does so mainly by driving cus-

tomers away until demand falls enough to reduce work pressure and stop the erosion of service quality (Figure 18). During temporary periods of high demand, the resulting drop in quality drives some customers away and causes some erosion of the service standard, forcing the workforce below initial levels. Temporary periods of low demand increase quality, and can win back some customers, but due to the asymmetry in standard adjustment, lags in adjusting the workforce, and other nonlinearities (e.g. in the customers' response to quality), service attractiveness spends more time below normal than above, resulting in persistent erosion of the customer base and a subsequent reduction of the workforce.

Policy recommendations

The simulation results show that service organizations are vulnerable to a wide range of self-reinforcing processes that can act as death spirals. In each case, the short-term benefits of an action, whether overtime, corner-cutting, service standard erosion or hiring, can trigger harmful long-term effects that either lead to insufficient capacity, drive away customers, or reduce the organization's budget, forcing further service capacity erosion. Although in principle these positive feedbacks can act as virtuous cycles, progressively improving service capability, leading to higher standards, more customers and revenue, and still greater service capacity, in practice, and as verified by empirical studies, the system is biased towards quality erosion. Workers are typically more willing to cut corners than increase their work effort, norms for the time that should be devoted to each customer fall more readily than they rise, management is more willing to raise its estimates of labor productivity than to cut them and to downsize rather than hire. Policies to reduce the strength and likelihood of quality erosion death spirals must overcome each of these processes.

Expediting capacity acquisition: Because 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 having a more responsive hiring process or reducing the delays governing capacity acquisition. Other strategies to increase the responsiveness of service capacity include hiring employees with greater initial effectiveness, accelerate the learning process by task routinization, maintaining a contingent workforce that can be deployed quickly when demand surges, and coordinating capacity management with other actions that affect demand such as marketing campaigns and product promotions. Unfortunately, these options are rarely available in high-contact services that require job-specific knowledge.

Reducing the effect of work pressure on time per task: Reducing employees' willingness to cut corners should slow the decline of service standards. Of course, if the time spent with customers were completely unaffected by work pressure, there could be no quality erosion. Such a rigid policy is unrealistic, however, because customer needs differ, individual servers have considerable autonomy in selecting how they respond to each customer, and the overtime required to hit

throughput targets with no flexibility in service would be prohibitive. A more realistic policy is to distribute employee responses to work pressure more evenly between corner cutting and overtime, while still responding fully to changes in work pressure. This can be done by reducing the flexibility of the service encounter (through process standardization and documentation, such as checklists in medical care) or by increasing the relative attractiveness of overtime (by creating high empathy with customers or increasing overtime compensation).

Creating quality pressure: Because service quality is intrinsically subjective and difficult to measure, there is little pressure from quality norms to counteract cuts in service induced by high work pressure. Surveys of customer experience are infrequent, less salient, and appear to be less consequential than the throughput, cost, and productivity feedback workers and service center managers receive every day. Though service providers often report some discomfort with their performance, we found no evidence in our fieldwork that low quality had any impact on the time employees devoted to each customer. Creating quality pressure requires management to become aware of the implications of poor service—lost sales, rework, and customer defections—and then, through training, incentives, measurement, and example, persuade employees that avoiding these costs is a priority and that they will not be punished for slowing their work to correct any quality problems they detect.

Maintaining a reserve margin of capacity: Even if the policies above are instituted, they will have little impact if the budget for service is continually tightened. Many of the executives responsible for customer service in organizations we have worked with continually face pressure to cut their expenses, even as the load on the service organization increases. It is common for managers at all levels, from supervisors to the CIO, to be told "technology is improving, and our shareholders expect double digit net income growth. You have to do more with less." The long delays in adjusting service capacity coupled with unpredictable variations in service demand mean an organization must maintain a strategic margin of reserve capacity to avoid the corner cutting, standard erosion, and other behaviors that trigger the death spirals. However, to many senior managers, reserve capacity looks like waste, leading to continual pressure to reduce budgets and headcount. Worse, financial stringency often prevents organizations from undertaking the process improvement initiatives that could lead to genuine improvements in productivity (Repenning and Sterman, 2001, 2002).

Figure 19 implements the policies suggested above. The hiring delay is cut from 20 to 10 weeks. Because corner cutting is not likely to be fully eliminated, we assume employees, through changes in training and incentives, are now twice as willing to work overtime as to cut corners. Quality norms are more resistant to erosion (the time for the standard to erode is lengthened from 20 to 25 weeks, still shorter than the 30 weeks for upward adjustment). Finally, a 5% margin of reserve capacity is built into the budget and staffing levels. Figure 19 shows the results of 1,000 monte-carlo simulations with different realizations of the random process for task arrivals.

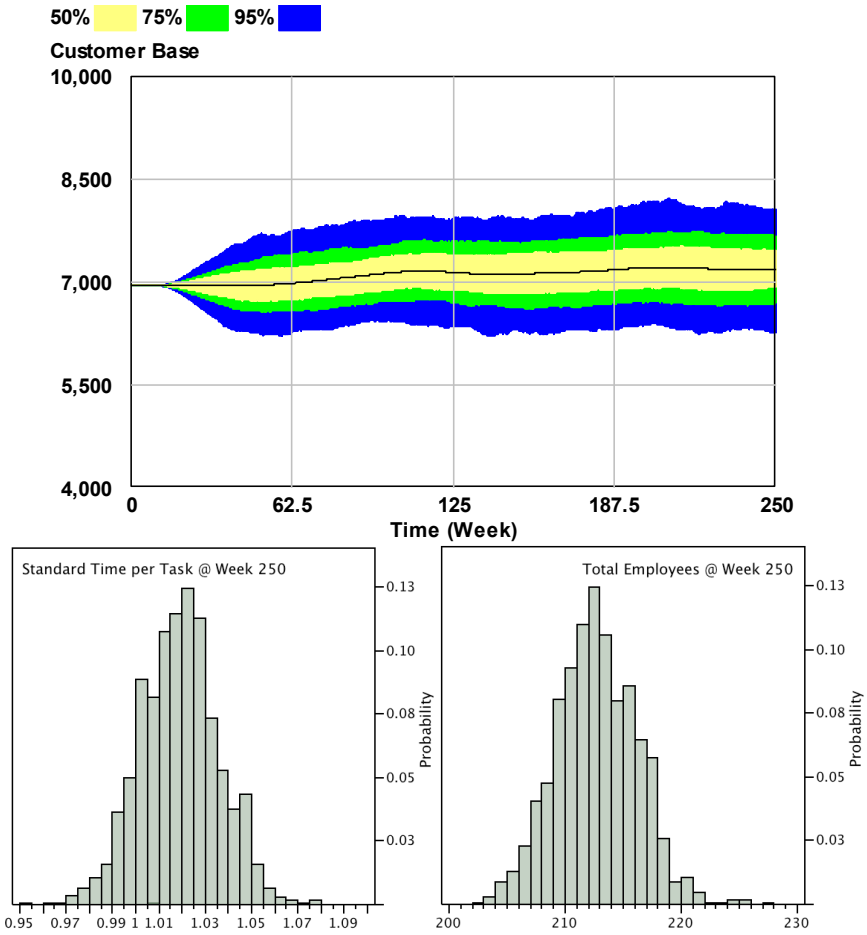


Figure 19. Response to stationary random orders with 10% standard deviation – Full model (policy recommendations).

On average the policies result in a rise in service standards and an increase in the customer base. The margin of reserve capacity lowers work pressure enough to enable servers to spend enough additional time with customers to push service quality up. Management is assumed to pay more attention to quality feedback, so the service standard also gradually increases. Higher average service quality gradually builds the customer base, and with it, the budget. The margin of reserve capacity also means that the growth of the customer base and therefore the task arrival rate does not immediately lead to high work pressure despite the lags in building service capacity. When random variations in task arrivals push work pressure above normal, service quality does not fall as much as before because servers are less willing to cut corners. These policies reverse the vicious cycles

that previously create the potential for a self-reinforcing death spiral; now the same positive feedbacks operate as virtuous cycles, leading the organization to progressively higher quality, longer employee tenure, higher productivity and lower error rates, building the customer base and revenue.

Concluding remarks

Despite the growing importance of services and service quality as sources of competitive advantage, the quality of service delivery in the United States is not improving and in many sectors is falling. Poor service quality contrasts sharply with the generally rising quality of manufactured products. Service delivery is intrinsically dynamic, involves multiple feedbacks among servers, customers, managers, and other actors. In this paper we develop a dynamic, behavioral model with a broad boundary to capture the structural characteristics of the service delivery process, management's and employees' decision-making processes, and the formation of expectations for customers and employees. Like any model, it is imperfect and can be extended; we provide full documentation and the model itself online so that others can replicate, extend, and improve it. Additional scenarios can be tested easily and tend to strengthen the results shown here. In particular, the propensity for capacity to become inadequate, triggering the positive feedbacks identified here, is greater in the presence of absenteeism and, especially, rapid growth in demand (e.g., Oliva et al., 2003).

To develop intuition for the dynamics of the service delivery process, we built the model up in stages to highlight the dynamics of each major sector – human resources, work flow, standards for quality, hiring, customer reactions to quality, budgeting, etc. By gradually expanding the boundary of the model we identify the many self-reinforcing feedbacks that can lead to persistent undercapacity and quality erosion. These feedbacks operate at the level of individual servers' responses to work pressure, in the interactions between hiring and experience, in the routines management uses to assess productivity and set staffing levels, and in the interactions between service quality and customer retention. High work pressure leads to overtime that causes fatigue and burnout, increasing errors and rework, lowering productivity, and increasing absenteeism and attrition, all of which feed back to worsen work pressure. High work pressure leads to corner cutting by employees, but the improvement in processing rates is interpreted by management as productivity growth, leading to staff cutbacks that raise work pressure further. Low quality causes customers to defect, cutting revenues and forcing further staff reductions. And so on.

Through simulations of the model we demonstrate that major recurring problems observed in the service industry—erosion of service quality, high turnover, and low profitability—can be explained by the organization's internal responses to work pressure. The manner in which a service firm responds to work pressure determines whether the system will disappoint customers, employees, and shareholders. Although in principle the positive feedbacks can operate as virtuous as

well as vicious cycles, the system is biased toward quality erosion by basic asymmetries and nonlinearities. Workers are typically more willing and able to cut corners (reduce the time spent with customers, cut back on preparation, documentation and other procedures to ensure fairness and quality) than to work overtime to maintain quality. Norms and standards for quality erode more easily than they rise. Because quality is subjective and difficult to measure compared to the relentless pressure to hit cost and throughput targets, management tends to interpret improvements in the customers processed per server as signs of productivity improvement even when they arise from corner cutting that harms the customer experience and erodes revenue. Hiring and building a skilled workforce is slower and more difficult than laying people off and losing skilled and motivated workers through burnout and voluntary attrition. Budgets are cut more readily than raised. And so on. These asymmetries mean service organizations are more likely to tip into death spirals in which the many positive feedbacks in the system operate as vicious cycles than to experience self-reinforcing improvement.

However, we also present policies that can reverse the death spiral and convert the vicious cycles into virtuous cycles of continuous improvement. Although quality, costs, and employee satisfaction are normally perceived as tradeoffs, we find successful policies that simultaneously delight customers, employees, and shareholders.

References

Aho, K. (2008). The customer service hall of shame. Retrieved 3/24/08, from http://articles.moneycentral.msn.com/SmartSpending/ConsumerActionGuide/TheCustomerServiceHallOfShame.aspx

Argote, L., & Epple, D. (1990). Learning curves in manufacturing. *Science, 247*, 920-924.

Baumol, W., Blackman, S. B., & Wolf, E. (1991). *Productivity and American Leadership*. Cambridge, MA: MIT Press.

Chesbrough, H. (2005). Toward a science of service. *Harvard Business Review, 83*, 16-17.

Chesbrough, H., & Spohrer, J. (2006). A research manifesto for services science. *Communications of the ACM, 49*(7), 35-40.

Darr, E., Argote, L., & Epple, D. (1995). The acquisition, transfer and depreciation of knowledge in service organizations: Productivity in franchises. *Management Science, 41*(11), 1750-1762.

Dogan, G. (2007). Bootstrapping for confidence interval estimation and hypothesis testing for parameters of system dynamics models. *System Dynamics Review, 23*(4).

Homer, J. B. (1985). Worker Burnout: A Dynamic Model with Implications for Prevention and Control. *System Dynamics Review, 1*(1), 42-62.

Horn, P. (2005, January 21, 2005). The new discipline of services science. *Businessweek*.

Jarmain, W. E. (Ed.). (1963). *Problems in Industrial Dynamics*. Cambridge, MA: MIT Press.

Lant, T. K. (1992). Aspiration Level Adaptation: An Empirical Exploration. *Management Science, 38*(5), 623-644.

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of Aspiration. In J. M. Hunt (Ed.), *Personality and the Behavior Disorders* (pp. 333-378). New York: The Ronald Press Company.

Lyneis, J. M., & Ford, D. N. (2007). System dynamics applied to project management: a survey, assessment, and directions for future research. *System Dynamics Review, 23*.

Maglio, P. P., Kreulen, J., Srinivasan, S., & Spohrer, J. (2006). Service systems, service scientists, SSME, and innovation. *Communications of the ACM, 49*(7), 81-85.

McGregor, J., McConnon, A., & Kiley, D. (2009, March 2, 2009). Customer service in a shrinking economy. *BusinessWeek*.

Oliva, R. (2001). Tradeoffs in responses to work pressure in the service industry. *California Management Review, 43*(4), 26-43.

Oliva, R. (2002). *Southwest Airlines in Baltimore (TN)* (Teaching Note No. 603-055). Boston, MA: Harvard Business School.

Oliva, R., & Bean, M. (2008). Developing operational understanding of service quality through a simulation environment. *International Journal of Service Industry Management, 19*(2), 160-175.

Oliva, R., & Sterman, J. D. (2001). Cutting corners and working overtime: Quality erosion in the service industry. *Management Science, 47*(7), 894-914.

Oliva, R., Sterman, J. D., & Giese, M. (2003). Limits to growth in the new economy: Exploring the 'get-big-fast' strategy in e-commerce. *System Dynamics Review, 19*(2), 83-117.

Repenning, N. P., & Sterman, J. D. (2001). Nobody ever gets credit for fixing problems that never happened. *California Management Review, 43*(4), 64-88.

Repenning, N. P., & Sterman, J. D. (2002). Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly*, 265-295.

Schneider, B. (1991). Service Quality and Profits: Can you have your cake and eat it, too? *Human Resource Planning, 14*(2), 151-157.

Schneider, B., Parkington, J. J., & Buxton, V. M. (1980). Employee and Customer Perceptions of Service in Banks. *Administrative Science Quarterly, 25*(2), 252-267.

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Boston: Irwin McGraw-Hill.

Strandvik, T. (1994). *Tolerance zones in perceived service quality*. Helsinki: Svenska handelshögskolan.

Thomas, H. R. (1993). *Effects of Scheduled Overtime on Labor Productivity: A Literature Review and Analysis* (Source Document No. 60). University Park, PA: Pennsylvania State University.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1993). The nature and determinants of customer expectations of service. *Journal of the Academy of Marketing Science, 21*(1), 1-12.