Cutting Corners and Working Overtime: Quality Erosion in the Service Industry

Rogelio Oliva • John D. Sterman
Harvard Business School, Morgan Hall T87, Boston, Massachusetts 02163
MIT Sloan School of Management, 30 Wadsworth Street, E53-351, Cambridge, Massachusetts 02142
roliva@hbs.edu • jsterman@mit.edu

The erosion of service quality throughout the economy is a frequent concern in the popular press. The American Customer Satisfaction Index for services fell in 2000 to 69.4%, down 5 percentage points from 1994. We hypothesize that the characteristics of services—inseparability, intangibility, and labor intensity—interact with management practices to bias service providers toward reducing the level of service they deliver, often locking entire industries into a vicious cycle of eroding service standards. To explore this proposition we develop a formal model that integrates the structural elements of service delivery. We use econometric estimation, interviews, observations, and archival data to calibrate the model for a consumer-lending service center in a major bank in the United Kingdom. We find that temporary imbalances between service capacity and demand interact with decision rules for effort allocation, capacity management, overtime, and quality aspirations to yield permanent erosion of the service standards and loss of revenue. We explore policies to improve performance and implications for organizational design in the service sector.
(*Organizational Learning; Service Management Performance; Service Operations; Service Quality; Simulation; System Dynamics*)

1. Introduction

Over the last decade, demand for customization has forced manufacturers to bundle more services with their products and service providers to rely more on personal interactions between customers and employees (McKinsey Global Institute 1992). As services require more customer contact and customization—a shift toward "high-contact" services (Chase 1981)—the challenges facing service managers have grown beyond the operational tasks of balancing supply and demand and ensuring quality in an environment where consumption and production are inseparable.

First, service organizations generate value through the delivery of an intangible, and intangible services are difficult to describe to new customers. It is likewise difficult for customers to express precisely what they expect from the service. Because there is no agreed objective standard about the service to be delivered, the only criteria available to evaluate service quality are subjective comparisons of customers' expectations to their perception of the actual service delivered (Zeithaml et al. 1990). Further, customers do not evaluate service quality solely in terms of the outcome of the interaction; they also consider the process of service delivery. Service quality, a multidimensional construct encompassing all aspects of service delivery, is difficult to assess and communicate.

Second, services are typically produced in the presence of the customer, and customers often participate in the production process. The simultaneous provision and consumption of services bring employees and customers physically, organizationally, and psychologically close, blurring the boundary between employees and consumers and enabling each to influence the other's perceptions and expectations. Studies

0025-1909/01/4707/0894$5.00
1526-5501 electronic ISSN

show a positive relationship between the perceptions, attitudes, and intentions of employees and customers (Schneider et al. 1980, Tornow and Wiley 1991). The lack of objective and fixed service standards and the mutual influence between servers and consumers point to a coevolution of their perceptions and expectations.

Finally, the high degree of customization created by the personal interaction of customers and service providers means that significant productivity gains through capital substitution in high-contact services are difficult. Baumol (1967, Baumol et al. 1991) demonstrated that the unbalanced growth of productivity in two industries causes unit costs in the stagnant sector to grow persistently and cumulatively relative to that of the progressive sector. Increasing unit cost translates into financial pressure on firms in the stagnant sector.

1.1. Erosion of Service Quality

The challenges described above are well documented. Little work, however, has been done to understand the effects of these driving forces acting simultaneously in a service setting. We hypothesize that these characteristics often bias service centers to reduce—albeit unintentionally—the level of service they provide to their customers, and can lock them into a vicious cycle of eroding service quality. We first observed this phenomenon in the context of the insurance industry (Senge 1990, Senge and Sterman 1992). The hypothesis can be articulated as follows: Because of rising financial pressure driven by slow productivity growth, managers attempt to maximize throughput per employee and minimize expense ratios. Because it is relatively difficult to obtain productivity gains in high-contact services, maximizing throughput drives the employees to work harder and, eventually, to reduce the attention given to customers. In the absence of accurate assessments of service quality and customer satisfaction, managers construe the reduction of attention given to customers as productivity gains, and, consistent with their objective of minimizing cost, reduce their estimates of required service capacity. The consequences of reducing attention to customers—high costs of poor quality (e.g., rework), low customer loyalty, and high turnover of service personnel—while difficult to perceive, reduce financial performance, creating financial pressure that encourages further cost containment.

Underinvestment in service capacity is frequently masked by eroding operating standards, so that servers, their managers, *and* customers all come to expect mediocre service and justify current performance based on past performance. Because firms monitor and benchmark on each other's performance, industry norms reinforcing expense control and productivity become increasingly influential in shaping individual firm decisions, and entire industries become locked into a vicious cycle of underinvestment and standard erosion. Industrywide erosion of service quality has been frequently cited in the popular press (e.g., *Quality* 1998, Koepp 1987) and recently reported by the American Customer Satisfaction Index. The 2000 ACSI for services fell to 69.4%, down 5 percentage points from its 1994 value (American Society for Quality 2001).

How does an organization gradually slip into eroding service standards? More important, how can it get out of the trap? This paper explores the consequences of the interactions among the structural characteristics of service processes to seek insight into the dynamics of service quality. The paper follows in the tradition of research in organizational learning and adaptation showing how organizational behavior arises from the interactions of physical and institutional structures with boundedly rational decision making, often leading to unintended and dysfunctional outcomes (e.g., Barnett and Hansen 1996, Forrester 1961, Levinthal and March 1981, March 1991, Masuch 1985, Sastry 1997, Sterman et al. 1997). We go beyond most existing studies, however, by developing a formal model that is tightly grounded in and tested against a detailed field study, and that provides a tool to design and test policies to avoid or reverse the undesirable outcomes generated by existing structures and routines. The paper follows our research approach. First, we developed a formal model that integrates the structural elements of service settings (§2). We tested the model empirically through calibration to a research site—a consumer-lending service center in a major U.K. bank (§3). We then used the model to understand the sources and

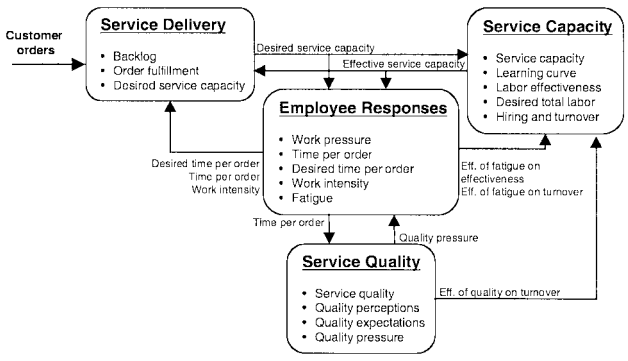
implications of service-quality erosion (§4) and generate some policy recommendations (§5). Finally, we discuss the implications of our findings for organizational theory and the service industry in general, and identify future research areas.

2. Model Structure

In this section, we present a formal model that integrates the characteristics of "high-contact" service. The model allows us to test whether service-quality erosion can be explained from structural elements of the service-delivery process—physical flows, organizational structure, and decision making—as opposed to variations unique to particular settings. Theoretical foundations and evidence for the hypothesized causal relationships are presented with each model equation.

The model consists of four sectors (Figure 1). The *service delivery* sector tracks the flows of customer orders through the service center. Service demand and standards determine the required service capacity. The *service capacity* sector models management's policies for setting staffing levels and renders a detailed account of hiring, on-the-job training, and turnover of the labor force. The *employee responses* sector models the way employees deal with the inevitable imbalances between demand and capacity by adjusting work hours and the time allocated to each customer. Finally, the *service quality* sector tracks the perception and formation of expectations of service quality for three types of agents in the service center—customers, employees, and managers—and models the impact of perceived quality on service operations.

Figure 1 Model Structure Overview



Service Delivery. The service-delivery sector tracks customer orders as they flow through the service center and determines the service capacity required to process the orders under current service standards. Customer orders (s_o) accumulate in a backlog (B) until they are processed. The order rate is exogenous. Exogenous orders imply that customers do not know the size of the backlog and cannot easily balk or renege after they enter the system—consistent with service operations such as insurance claims and banking. The backlog is reduced by the order-fulfillment rate (s_f),

$$(d/dt)B = s_o - s_f. \tag{1}$$

The order-fulfillment rate (s_f) is effective service capacity (c) adjusted by the employees' work intensity (i)—the fraction of time available allocated to processing orders—and divided by the actual time allocated to fulfill a customer order (T). In the case of excess capacity, the order-fulfillment rate is limited by the orders that can be processed from the backlog and the minimum time required to process orders (τ_f),

$$s_f = \min(c \cdot i/T, B/\tau_f). \tag{2}$$

Required service capacity (c^*) is given by the backlog of unfulfilled orders (B), management's goal for delivery delay (λ), and the standard for the time to be allocated to each customer (T^*),

$$c^* = (B/\lambda) \cdot T^*. \tag{3}$$

Service Capacity. The service-capacity sector models hiring, on-the-job training, and turnover of the labor force.[1] Not all employees have the skills and/or energy required to perform the job with the same productivity, hence the traditional definition of service capacity—time available for processing orders—is expanded to include effects of worker skill and effort. Effective service capacity (c) is determined by

[1] The original formulation of the model (Oliva 1996) included a CES production function with capital stocks and their technological content. However, for most ranges of reasonable parameters, including those of the research site, the dynamics of capital substitution proved to be much slower than the dynamics described in this paper, hence, here capital is assumed constant.

adjusting the total labor force (L) by the effects of personnel experience (e) and fatigue (f) (Equations (6) and (27)),

$$c = L \cdot e \cdot f. \tag{4}$$

Learning-by-doing is well documented in a wide range of settings, including service-delivery organizations (Argote and Epple 1990, Darr et al. 1995). The importance of customization suggests potential for significant learning in high-contact service settings, and, indeed, our fieldwork found evidence of such learning. When services involve personal and customized interaction between individual servers and customers, much of the learning gained through experience will be embodied in the skills and behaviors of the individual workers. We model the individual learning curve of new employees as an "experience chain" (Jarmain 1963). New hires are assumed to have only a fraction (ε) of the productivity of more-experienced employees, but through on-the-job coaching, mentoring, and experience, they gradually gain skills that boost their productivity. Mentoring and on-the-job coaching are not free—each new hire reduces the productivity of experienced personnel by a constant fraction (η) during the training period. Labor (L) is separated into two populations: experienced personnel (L_e) and rookies (L_r). The mix of the two populations and their relative productivity determine the effect of personnel experience (e), which affects service capacity (Equation (4)). The effect of experience is the number of full-time-equivalent experienced personnel relative to the total labor force,[2]

$$L = L_e + L_r, \tag{5}$$

$$e = \max(0, (L_e + L_r(\varepsilon - \eta))/L) \quad 0 \le \varepsilon \le 1, \eta \ge 0. \tag{6}$$

Equations (7)–(11) account for the flow of employees through the experience chain and on-the-job learning. The stock of rookies is increased by the hiring rate (l_h) and decreased as employees become experienced (l_e). The stock of experienced personnel is augmented as rookies gain experience (l_e) and reduced by attrition (l_a). The experience rate (l_e) captures the transition from rookies to experienced personnel. Rookies develop full productivity through a first-order process characterized by an average training period (τ_e), a proxy for cumulative experience,[3]

$$(d/dt)L_r = l_h - l_e, \tag{7}$$

$$(d/dt)L_e = l_e - l_a, \tag{8}$$

$$l_e = L_r/\tau_e. \tag{9}$$

Turnover from the experienced-personnel stock is assumed to be exponential with an average time for turnover (τ_a). The training period is relatively short compared with the average tenure of employment; hence, we ignore turnover from the rookie stock. Attrition depends on factors external and internal to the firm, including the health of the economy and labor market, organizational attributes, and worker-specific factors (Mobley 1982). The economic factors are considered exogenous to the model and captured in the nominal turnover time (τ_a^*). Two organizational attributes are modeled endogenously and modify the nominal turnover time: employees' fatigue (a_f) and perception of service quality (a_q); high fatigue and low quality both lead to more turnover (Equations (29) and (34)):

$$l_a = L_e/\tau_a, \tag{10}$$

$$\tau_a = \tau_a^* \cdot a_f \cdot a_q. \tag{11}$$

It takes time to hire new employees. Equations (12)–(17) portray the labor supply chain (unfilled vacancies) and the hiring policies as a stock-management problem (Sterman 1989). The hiring rate depends on the firm's unfilled labor vacancies (L_v) and a hiring delay (τ_h). Vacancies represent the labor

[2] The effective labor fraction (e) is constrained to be nonnegative to control for cases where rookies require more supervision than their initial effectiveness ($\eta \gg \varepsilon$) and rookies outnumber the senior personnel ($L_r \gg L_e$).

[3] The experience chain represents learning as human capital embodied in individual workers, and differs from the traditional formulation in which learning is a function of cumulative experience. The two formulations are related because individual workers accumulate experience at a constant rate (1 week/week). Zangwill and Kantor (1998) examine the relationships among different formulations for learning; see also Argote and Epple (1990).

orders (l_o) that have not been filled. By Little's law, desired vacancies (L_v^*) are proportional to the desired hiring rate and the hiring delay (τ_h), the time it normally takes to fill a vacancy,

$$l_h = L_v/\tau_h, \tag{12}$$

$$(d/dt)L_v = l_o - l_h, \tag{13}$$

$$L_v^* = l_h^* \cdot \tau_h. \tag{14}$$

Indicated labor orders (l_o^*) are determined by the desired hiring rate (l_h^*) corrected for any discrepancies between desired and actual vacancies ($L_v^* - L_v$). Similarly, the desired hiring rate is determined by the replacement of employees that have departed the service center (l_r) (except when trying to downsize), corrected for any discrepancy between desired and existing labor ($L^* - L$). The responsiveness of the policy to close each of these gaps is given by the time to adjust labor (τ_l),

$$l_o^* = l_h^* + (L_v^* - L_v)/\tau_l, \tag{15}$$

$$l_h^* = l_r + (L^* - L)/\tau_l \quad l_r = \begin{cases} 0 & \text{if } L > L^* \\ l_a & \text{otherwise.} \end{cases} \tag{16}$$

If indicated labor orders are negative, the order rate is limited to the number of unfilled vacancies that can be canceled and the time it takes to do so (τ_v),

$$l_o = \max(-L_v/\tau_v, l_o^*). \tag{17}$$

Finally, the desired number of employees (L^*) is determined from management's perception of labor effectiveness (E) and required service capacity (c^*). We assume, a fortiori, that hiring is not constrained by financial considerations that often cause underinvestment in service capacity. Instantaneous labor effectiveness, defined by effective service capacity per worker (c/L), is not immediately perceived. Management's perception of labor effectiveness (E) is assumed to be perceived after a delay (τ_{pe}) representing the time required to measure, report, and assess changes in productivity. Because labor is costly and slow to change, management does not act on instantaneous labor requirements (c^*/E). Instead, desired labor (L^*) adjusts by exponential smoothing with time constant (τ_l^*) to filter out high-frequency noise in demand,

$$(d/dt)E = ((c/L) - E)/\tau_{pe}, \tag{18}$$

$$(d/dt)L^* = ((c^*/E) - L^*)/\tau_l^*. \tag{19}$$

Employee Responses. Delays in adjusting service capacity and the variability of customer orders make it extremely difficult to balance supply and demand in an environment where service delivery and consumption are simultaneous. Work pressure (w), a measure of the balance between service demand and capacity, is defined as the gap between required service capacity and effective service capacity as a fraction of current capacity,

$$w = (c^* - c)/c. \tag{20}$$

Work pressure can also be interpreted as the relative workload in the service center. Employees respond to work pressure by adjusting their behavior to meet throughput expectations. The first response to a change in work pressure is for employees to adjust the time allocated to each order (T). An anchoring and adjustment process (Einhorn and Hogarth 1981) is assumed. Employees select a service level by anchoring on the current service standard, then adjusting actual service above or below the standard in response to the current workload (t_w) and quality pressure (t_p). In turn, the level of service actually delivered modifies the anchor (Hogarth 1980). Because a given absolute difference between desired and actual performance becomes psychologically less important as actual performance increases, the adjustment process is multiplicative (Kahneman and Tversky 1982). The formulation constitutes a hill-climbing search process that does not require knowledge of the function linking the amount of time dedicated per customer order to delivered quality—an assumption consistent with the intangibility of service quality. The search process is limited by the minimum amount of time required to process a customer order (τ_f),

$$T = \max(t_w \cdot t_p \cdot T^*, \tau_f). \tag{21}$$

The effects of work pressure and quality pressure—the normalized gap between employees' perception

of delivered service quality and their quality expectation—on time per order (t_w and t_p) are assumed to be nonlinear and to be neutral in the absence of pressure,

$$t_w = f_{wt}(w) \quad f(0) = 1,\ f' \le 0, \tag{22}$$

$$t_p = f_{pt}(p) \quad f(0) = 1,\ f' \ge 0. \tag{23}$$

The adjustment process for the underlying standard for time per order, the time employees would allocate to each order in the absence of work and quality pressure, is asymmetric. Asymmetric adjustment processes have been used in the organizational and psychological literature to represent the biased formation of expectations and goals (Lant 1992), and are normally formulated by allowing different time constants to govern the adjustment process, depending on whether the aspiration level is above or below actual performance,

$$(d/dt)T^* = (T - T^*)/\tau_{to} \quad \tau_{to} = \begin{cases} \tau_{ti} & \text{if } T > T^* \\ \tau_{td} & \text{otherwise.} \end{cases} \tag{24}$$

The second way employees deal with high work pressure is by increasing their work intensity by taking shorter breaks or working overtime. In the model, employees adjust work intensity (i) in response to work pressure (w). The response is nonlinear, and limited by the time an employee could be working,

$$i = f_{wi}(w) \qquad f(0) = 1,\ f(\infty) = i^{\max},\ 0 \le f' \le 1. \tag{25}$$

Extended periods of high work intensity, however, cause fatigue that eventually undermines the productivity gains achieved through longer hours (Homer 1985, Thomas 1993). In the model, fatigue (F_e) is captured by exponential smoothing of work intensity (i) over the average time required for fatigue to set in (τ_{fe}). The effect of fatigue on effectiveness (f) is a decreasing nonlinear function that reduces effective service capacity when service personnel are tired (Equation (4)),

$$(d/dt)F_e = (i - F_e)/\tau_{fe}, \tag{26}$$

$$f = f_{fe}(F_e) \quad f(F_e \le 1) = 1,\ f' \le 0,\ f'' > 0. \tag{27}$$

Extended periods of high work intensity also have an impact on average employee tenure (Farber 1983, Mobley 1982, Weisberg 1994). A formulation similar to the effect of fatigue on productivity is used to capture the effect of fatigue on employee attrition (a_f; Equation (11)). The time constant for the fatigue level driving attrition is τ_{fa}. While extended overtime quickly affects productivity, the impact of burnout on attrition is slower; hence, $\tau_{fa} > \tau_{fe}$,

$$(d/dt)F_a = (i - F_a)/\tau_{fa}, \tag{28}$$

$$a_f = f_{fa}(F_a) \quad f(F_a \le x) = 1,\ f(\infty) = 0,\ f' \le 0. \tag{29}$$

Service Quality. To address the issues of service *inseparability* and *intangibility*, we define service quality as a function of customers' expectations and the time allocated per customer. Because time per order adjusts to changes in effective labor capacity, it functions as a proxy for the degree of attention and care that servers are providing. Perceived service quality suffers if customers feel rushed by the servers, or perceive a poor attitude or lack of skills. As more effective time is allocated to each order, employees are able to inquire into and satisfy customer needs beyond minimal transactional requirements. The assumption that time per order is the main driver of service quality is consistent with Mills's (1986) equation of service quality with server productivity and the common claim that "the most important component of a service is personnel" (Broh 1982). The metric also captures four of the five dimensions of service quality identified by Zeithaml et al. (1990)—reliability, responsiveness, assurance, and empathy.

Customer expectations are modeled as customers' beliefs regarding the *effective* time that should be allocated to each order (T_c^*). The satisfaction or quality customers experience (q) is a nonlinear function of the performance gap—the normalized difference between the time allocated per order (T) and customers' expectations (Zeithaml et al. 1990),

$$\begin{aligned} q &= f_q((T - T_c^*)/T_c^*) \\ f(0) &= 1,\ 0 \le f\{\cdot\} \le f^{\max},\ f' \ge 0. \end{aligned} \tag{30}$$

Although the exact relationship between effective time per order and service quality might vary from setting to setting, some generic characteristics can be specified. Experienced quality is one (acceptable)

when the time allocated to each customer equals the time they expect to be allocated. If the time allocated falls below the time expected, quality drops (to a minimum of zero). The existence of a "tolerance zone" for service quality (Strandvik 1994, Zeithaml et al. 1993) suggests a function that is relatively flat when $T \approx T_c^*$, but grows progressively steeper as the performance gap rises. Kano's differentiation of quality attributes between *must-be's* and *delighters* (Shiba et al. 1993) indicates that there are diminishing returns to the perceived value of an attribute, suggesting a saturation effect as performance rises above expectations.

The intrinsic subjectivity of quality means it takes time to perceive, measure, and report quality, and changes in customers' experiences will only be perceived by workers and management after a delay. The quality levels perceived by employees (Q_e), management (Q_m), and customers (Q_c) adjust via first-order exponential smoothing of actual quality. The time constants for these perceptual processes are assumed to be different, and ranked according to their immediacy to the delivery process and the frequency of exposure to it,

$$(d/dt)Q_g = (q - Q_g)/\tau_{qg} \text{ where } g \in \{e, m, c\}. \quad (31)$$

In addition to their perceptions of service quality, each agent involved in the service-delivery process—employees and customers—is assumed to have an internal standard for the service level that ought to be delivered. These expectations are conceptualized as levels of aspiration (Lant 1992, Simon 1957), and are modeled as a weighted average of prior aspiration level and perceptions of current performance (Cyert and March 1963, Levinthal and March 1981, Morecroft 1985). Because assessments of service quality are based on the gap between perceptions and expectations, the aspiration-adjustment process is particularly appropriate in the creation of quality expectations (Boulding et al. 1993).

Customers' expectations for how much time servers should spend with them are anchored to the service provided by competitors (μ) and adapt to the current service experienced (T_c),

$$(d/dt)T_c^* = (\omega_c \mu + (1 - \omega_c)T - T_c^*)/\tau_{ec}$$
$$0 \le \omega_c \le 1. \quad (32)$$

The employees' quality standard (Q_e^*) is assumed to adapt via exponential smoothing to a weighted average of the employee's own perception of the quality of service delivered to the customer (Q_e) and management's desired quality goals (Q_m^*),

$$(d/dt)Q_e^* = (\omega_e Q_e + (1 - \omega_e)Q_m^* - Q_e^*)/\tau_{ee}$$
$$0 \le \omega_e \le 1. \quad (33)$$

Perceptions and expectations of service quality feed back to the service-delivery process in two ways. First, the human resources literature shows that 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). Thus, when employees perceive quality is low, the average duration of employment falls (Equation (11)),

$$a_q = f_{qa}(Q_e) \quad f(0) = 0, f(1) = 1, f' \ge 0. \quad (34)$$

Second, the gap between employees' perceptions of delivered service quality (Q_e) and their quality expectations (Q_e^*) affects the time allocated per order. The dissonance created by this gap is defined as quality pressure (p) and is formulated analogously to work pressure (Equation (20)),

$$p = (Q_e^* - Q_e)/Q_e^*. \quad (35)$$

Because service quality is inseparable from the delivery process, and therefore the attitudes and behavior of the employees, changes in quality are driven by the gap between employee perceptions of quality and their aspirations ($Q_e^* - Q_e$). Management affects service quality indirectly, through changes in the employees' goals for service quality (Equation (33)).

3. Empirical Testing

Although the proposed model describes relationships that have been documented in the literature, much of the evidence available for those relationships is fragmented and case-specific; no full exploration of all the simultaneous interactions has been published. To test and build confidence in the model as a whole, it is

necessary to assess whether the individual relationships operate simultaneously in a wide range of service settings, and if their interactions are capable of replicating the observed behaviors of service settings (Forrester 1979, Naylor and Finger 1967, van Horn 1971). As a first step in this process, we tested the model against a particular service setting—a retail banking operation in the United Kingdom. We used data from this site to statistically estimate individual relationships in the model. We then compared the behavior of the full model against the available data, assessing the extent to which the model quantitatively replicates the observed behavior. We explored the robustness of the conclusions through sensitivity analysis and simulations of scenarios representing situations not experienced at the research site.

3.1. The Research Site

National Westminster Bank, Plc. is the flagship of NatWest Group, one of the largest financial institutions in the United Kingdom. In 1990, the U.K. Retail Banking Services (RBS) unit of NatWest sought to cut costs by moving back-office operations from branches to centralized processing centers in more affordable locations. Created in June 1993, the Lending Center (LC) at Nelson House serves as the back office for the mass market (personal loans and credit cards) and small business accounts (sales $\leq$ £100,000 per year) in the West End region of London. When our field work was done, the LC served 245,000 accounts distributed in 20 branches—about 2% of the total account volume of U.K. RBS—and had plans to integrate 11 additional branches over the next 18 months. In the LC, groups of lending officers are responsible for particular branches. Work arrives at the LC by phone (customer inquiries), mail (customer requests and communications with branches), and daily computer-generated reports identifying problematic accounts that require immediate action (such as overdrafts, missing payments, etc.). Most requests produce either a letter or a phone conversation with the customer. The variety of tasks performed is limited and order flows are monitored against standard processing times for each task type.

Data collected by the first author included (1) time series for key operational metrics; (2) interviews with employees, their managers, and staff, inside and outside the LC; (3) 12 hours of direct observation; and (4) archival data, such as policy and procedure manuals and training materials. We used these data to specify the decision rules of employees and managers. Wherever possible we used the numerical data to estimate parameters and relationships. Finally, from anecdotes and descriptions of unusual incidents we identified how the system responds to extreme conditions. Frequently, the different data-gathering methods allowed for triangulated measurements of the same relationship. The following subsection presents an example of model estimation for a critical decision—how much time employees allocate to each order—and the use of data from multiple sources to make sense of the statistical results. The remainder of the section summarizes the sources for parameter estimates and presents the model's fit to historical data.

3.2. Partial Model Estimation

We hypothesized (Equation (21)) that time per order (T) depends on the desired time per order (T^*), adjusted by the effects of work pressure (t_w) and quality pressure (t_p). The adjustment, however, does not occur in a vacuum. Time per order (T) and desired time per order (T^*) are tightly coupled through two feedback loops—the "anchoring and adjustment" process (Equations (21) and (24)), and the "goal adjustment" that occurs as desired time per order determines required service capacity (Equations (3), (20), (21), (22), and (24)). Since desired time per order is not directly observable, we estimated the parameters governing its adjustment together with the response to work pressure (w). The effect of work pressure on time per order (t_w) was specified by the exponential function $\exp(\alpha w)$; the parameter α controls the response of time per order to work pressure. A separate partial model estimation showed that the effect of quality pressure on time per order was not statistically significant. This result is consistent with the observation that the LC did not have market research instruments in place to monitor and report customer satisfaction. The effect of quality pressure on time per order (t_p) is assumed constant in this partial model estimation (Equation (23′)). The estimation

minimizes the sum of squared errors between simulated and actual time per order given the structure of the model and driven by the data for actual service capacity (SC) and customer orders (CO):

$$\underset{T_0^*,\alpha,\tau_{ti},\tau_{td}}{\text{Min}} \sum_{t=1}^{n} (T(t) - TPO(t))^2$$

subject to

$$T(t) = \max(t_p(t) \cdot t_w(t) \cdot T^*(t), \tau_f);\ \tau_f = 0.1 \quad (21')$$

$$T^*(t) = \int (T(t) - T^*(t))/\tau_{to} + T_0^*;$$

$$\tau_{to} = \begin{cases} \tau_{ti} & \text{if}(T(t) > T^*(t)) \\ \tau_{td} & \text{otherwise} \end{cases} \quad (24')$$

$$sc^*(t) = CO(t) \cdot T^*(t) \quad (3')$$

$$w(t) = (sc^*(t) - SC(t))/SC(t) \quad (20')$$

$$t_w(t) = \exp(\alpha w(t)) \quad (22')$$

$$t_p(t) = 1 \quad (23')$$

We derived the service-capacity data series from the number of employees corrected for absenteeism and adjusted for the effects of fatigue and experience.[4] Because the LC cleared the backlog of orders every day, customer orders proxy the desired fulfillment rate (B/λ; Equation (3′)). The observed time per order (TPO) was calculated from the time allocated to processing orders (total time + overtime − absenteeism − training) divided by the number of orders processed. All data series were available from the LC's weekly operating reports from June 1994 through May 1995. Table 1 shows the estimated values for the parameters, with 95% confidence intervals. All estimates have the correct signs and tight confidence bounds. The fit between the simulated series and the historical data is presented in Figure 2. The Theil inequality statistics describe the fraction of the mean square error between simulated and actual series due to unequal means (bias), unequal variances, and imperfect correlation (Theil 1966). Low bias and variance fractions indicate that the error is unsystematic (Sterman 1984).

The initial estimate for desired time per order is 1.08 person-hours, about 7% less than the stated goal (bank procedures called for one hour of preparation and breaks for every 6 hours processing orders, implying desired time per order of 1.17 person-hours). Interviews suggested that service personnel worked unreported overtime that accounted for most of the discrepancy and direct observation corrobo-

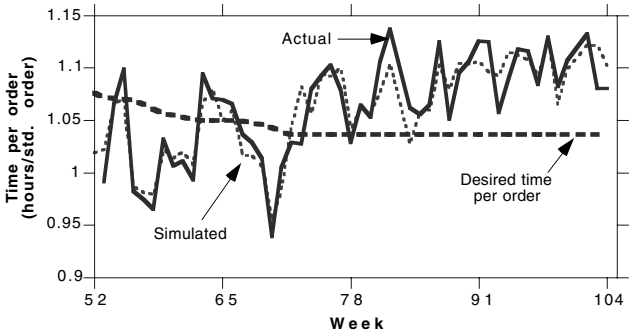
[4] Because the average work week in our dataset was always within 10% of the standard work week (35 hrs), the fatigue feedback was not active. Employee-experience mix and its effects on productivity were estimated independently with data from June 1993 to May 1995.

Table 1 Estimates for the Adjustment of Time per Order

	Estimate	95% Confidence Interval†	
T_0^*	1.08	1.06	1.09
α	−0.64	−0.70	−0.59
τ_{td}	18.83	13.30	28.95
τ_{ti}	814,000	327,000	∞

†Calculated from the curvature of the response surface without assumptions of symmetry.

Figure 2 Time per Order (Partial Model Estimation)



Summary Statistics for Historical Fit—Time per Order	
$n = 50$	
R^2	0.828
Mean Absolute Percent Error	1.5%
Root Mean Square Error	0.019%
Theil's Inequality Statistics	
Bias	0.000
Unequal Variation	0.047
Unequal Covariation	0.953

rated these statements:

> I don't claim it all in overtime. I tend not to claim for work I do before the eight o'clock start, nor for the lunch hour [approx. 5 hours/week].
>
> ... And they don't always claim that overtime either. I suppose that they're worried that someone would say "you are not working very clever" (sic) or something. I never go out to lunch; I'm giving the bank five hours a week of [unpaid] overtime.

The most important result of the partial model estimation is the asymmetry of the adjustment process for desired time per order. When work pressure forces actual time per order to fall below the desired level, the desired level erodes quickly, with an estimated time constant (τ_{td}) of about 19 weeks. But there is no evidence of any upward revisions in desired time per order when work pressure is low ($\tau_{ti} \approx \infty$), despite the fact that actual time per order exceeded desired time per order in more than half the dataset. High work pressure leads employees to reduce their aspirations for the time they should spend with each customer. But once they learn how to deliver the service faster, that ability and mindset seems to endure even in times of low work pressure.

3.3. Estimation Summary

Similar techniques were used to estimate parameters and initial conditions for the rest of the model. From data series of authorized labor, total labor, and hiring, it was possible to estimate the parameters of the service-capacity sector (Equations (7)–(17)). Parameters for management-staffing policies (Equations (3), (18)–(19)) were estimated from data on service capacity and authorized labor, and overtime reports were used to estimate the effect of work pressure on work intensity. Consistent with our hypothesis, management had no instruments in place to assess customer satisfaction operationally, thus the formation of quality standards was exclusively driven by employee perceptions of service quality ($\omega_e = 1$).[5] Once the formation of quality standards was identified, and assuming, a fortiori, constant customer-service expectations ($\omega_c = 1$), we used data on time per order and service capacity to estimate employee perceptions of service quality and the effects of quality pressure on time per order (Equations (31) and (23)). In the absence of time-series data, the parameters governing the employees' learning curve (τ_e, ε, and η) and their perceptions of and expectations for service quality (τ_{qe} and τ_{ee}) were selected based on interviews and surveys. Estimates of these parameters solicited from individual employees were quite consistent with one another.

Of 37 model parameters (including nonlinear functions and initial conditions), we estimated 14 econometrically and set another 5 directly from their historical values. We obtained 10 parameters through direct observation or interviews. Four parameters, all related to work intensity and its effects, were not active during the period for which data were available, and thus could not be estimated statistically. Although not active for simulations, we set these parameters to the best estimates available from the literature. Table 2 lists all parameters, their values, and sources.

3.4. Historical Fit of the Model

The derivation of model structure and parameters from the observed physical structure and decision rules, and the ability of partial model structure to replicate data series with plausible parameters, constitute tests of the model's structural validity (Barlas 1989, Forrester and Senge 1980). Furthermore, the policies estimated for the decision makers show that their behavior is locally or intendedly rational relative to the existing incentive system (Morecroft 1985). The ability of the model to replicate historical behavior constitutes another test. We simulated the full model under historical conditions driven by only two exogenous data series: customer orders and absenteeism. We assessed model behavior against six variables for which time series were available (Figure 3).

The mean absolute percent error (*MAPE*) between the simulated and actual variables is less than 2% for

[5] NatWest RBS did have an instrument to monitor quarterly customer satisfaction, but the questionnaire was designed with the traditional customer service branch in mind, thus the information collected was of little use. The LC collects monthly satisfaction surveys from the managers of the branches that it serves but, according to the LC management, the information was neither reliable nor useful.

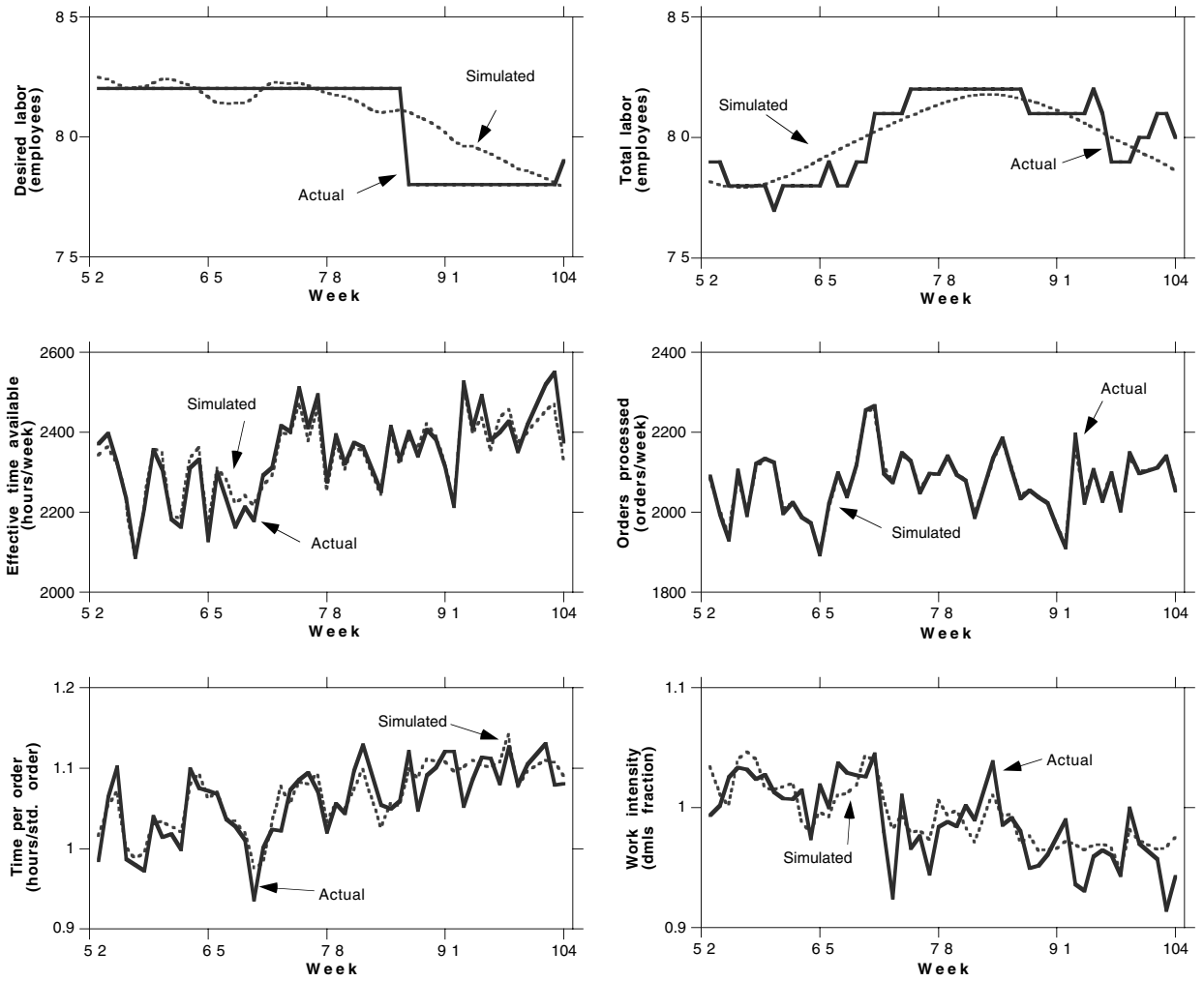
Table 2 Parameters and Sources for Service Model

Parameter		Value		Source
Service delivery				
τ_f	Minimum time required to process an order	0.1	week	Set based on observations
λ	Desired delivery delay	0.1	week	Set based on stated goals
Service capacity				
τ_l	Time to adjust labor	11.5	week	Estimated to fit past data on labor hiring
τ_h	Hiring delay	29.9	week	Estimated to fit past data on labor hiring
τ_a	Time for attrition	401.0	week	Estimated to fit past data on attrition
τ_v	Time to cancel vacancies	1.0	week	Set based on stated procedures
τ_{pe}	Time to perceive labor effectiveness	6.7	week	Estimated to fit past data on desired labor
τ_l^*	Time to adjust desired labor	18.8	week	Estimated to fit past data on desired labor
τ_e	Time for experience	12.0	week	Judgmentally set based on interviews
ε	Relative effectiveness of rookies	0.35	dimensionless	Judgmentally set based on interviews
η	Fraction of experienced personnel for training	0.05	dimensionless	Judgmentally set based on interviews
Employees' responses				
f_{wt}	Effect of workload on time per order	$e^{-0.64w}$	dimensionless	Estimated to fit past data on time per order
τ_{ti}	Time for upward adjustment of time per order	813, 564	week	Estimated to fit past data on time per order
τ_{td}	Time for downward adjustment of time per order	18.8	week	Estimated to fit past data on time per order
f_{wi}	Effect of workload on work intensity	$e^{0.37w}$	dimensionless	Estimated to fit past data on work intensity
τ_{fe}	Time for effect of fatigue on effectiveness	3.0	week	Set based on previous studies
τ_{fa}	Time for effect of fatigue on attrition	52.0	week	Set based on previous studies
f_{fe}	Effect of fatigue on effectiveness $F_e \in [1.14, 2]$	$1\text{-}0.5F_e$	dimensionless	Set based on previous studies
f_{fa}	Effect of fatigue on attrition $F_a \in [1, 2]$	$1\text{-}0.2F_a$	dimensionless	Set based on previous studies
Service quality				
ω_c	Weight for customers' service expectation	1.0	dimensionless	Set a fortiori and based on interviews
ω_e	Weight for employees' quality expectation	1.0	dimensionless	Set based on interviews
μ	Customers' service expectation reference	1.16	hours/order	Estimated to fit past data on time per order
f_{pt}	Effect of quality pressure on time per order	$e^{0.00p}$	dimensionless	Estimated to fit past data on time per order
f_{qa}	Effect of quality on attrition	1.00	dimensionless	Set based on historical data
τ_{qe}	Time for employees' perception of quality	4.0	week	Judgmentally set based on interviews
τ_{ee}	Time for employees' quality expectation	26.0	week	Judgmentally set based on interviews
Q_m^*	Management quality goal			Not active in base simulation
τ_{qm}	Time for management's perception of quality			Not active in base simulation
τ_{qc}	Time for customers' perception of quality			Not active in base simulation
τ_{ec}	Time for customers' service expectation			Not active in base simulation
Initial conditions†				
L_e	Experienced personnel	64.0	employees	Set based on historical data
L_r	Rookies	14.0	employees	Set based on historical data
E	Perception of labor effectiveness	0.78	dimensionless	Estimated to fit past data on desired labor
T^*	Desired time per order	1.08	hours/order	Estimated to fit past data on time per order
F_e	Fatigue for effect on employee effectiveness	1.00	dimensionless	Set based on historical data
F_a	Fatigue for effect on employee attrition	1.00	dimensionless	Set based on historical data
Q_e	Employees' perception of quality	0.95	dimensionless	Estimated to fit past data on time per order

†The rest of the stocks were initialized in equilibrium from known parameters.

all series (Table 3). The low bias and variation components of the Theil inequality statistics indicate that the errors are unsystematic. The model's exceptionally good tracking of orders processed arises because employees sought to process all orders each day and because overtime, time per order, and hiring varied enough to prevent capacity shortfalls. The relatively low R^2 in some of the comparisons is caused by the high-frequency noise in customer orders and absenteeism. The model functions as a low-pass filter

Figure 3 Comparison of Simulated and Actual Data



capable of tracking the overall behavior of the system variables, but it is not suitable for point predictions of random day-to-day events.

The simulation begins 52 weeks after the creation of the LC and runs for a year. During this period no additional branches were incorporated into the LC, and demand remains stationary (see orders processed in Figure 3). However, there is a substantial labor shortage during the first half year as the LC ramps up its staff. Employees compensate through overtime (work intensity is greater than one). Aggressive hiring during the first 6 months increases the time available to process orders, reducing work intensity. By Week 80, the labor deficit is closed and hiring slows. After Week 84, despite the fact that orders remain stationary, there is an overshoot in service capacity. Initial estimates of required labor were made under growth conditions, when a high fraction of the workers were inexperienced and required training. Once hiring slows, training requirements fall. As new employees gain experience, they become more productive and require less supervision, increasing the effective time available for order processing. Even though management updates its estimate of labor productivity, there is enough momentum in the system (from rookies gaining experience) to cause capacity to overshoot and work intensity to drop.

Table 3 Historical Fit June 1994–May 1995

		Theil's Inequality Statistics				
	MAPE	Bias	Unequal Variation	Unequal Covariation	R^2	*N*
Desired labor	0.9	0.109	0.257	0.633	0.740	52
Total labor	0.8	0.026	0.143	0.830	0.747	52
Time available	0.9	0.019	0.255	0.725	0.938	50
Orders processed	0.3	0.000	0.299	0.701	0.990	50
Time per order	1.7	0.033	0.095	0.872	0.799	50
Work intensity	1.7	0.060	0.154	0.784	0.635	50

4. Analysis

The disequilibrium in the historical case provides a good test of the model and our proposed hypothesis for erosion of service quality. First, the simulation fits the historical data quite well, thus increasing our confidence in the proposed model. Second, the historical simulation shows some evidence of erosion of the internal service standard—measured by desired time per order—during the first third of the simulated horizon (Figure 2). This erosion of the service standard, however, occurs when there is a labor shortage and when most employees are not fully experienced. To test the theory, we have to show that quality can erode during normal operations and not only during the transient as the LC initiates operations. To eliminate the transient effects of initial conditions, we tested the model in a stochastic equilibrium. The rest of this section presents a series of tests designed to isolate the structural characteristics contributing to quality erosion even when resources are, on average, in balance with demand.

4.1. Response to Historical Variations

We initialized the model in equilibrium with characteristics achieved by the LC after the transient ramp-up period shown in Figure 3. In Week 10 we introduce stochastic variations in customer orders and absenteeism. These were modeled as independent stationary random variables whose means, variances, and autocorrelation spectra were estimated from the historical data. Simulations of the equilibrium base case showed that employees absorb small increases in work pressure arising from variations in demand and absenteeism[6] by reducing time per order (the Corner Cutting Loop B1 in Figure 4) and increasing work intensity (the Overtime Loop B2). The reduction in time per order, while enabling an immediate increase in throughput, also erodes the internal service standard—desired time per order (the Goal Erosion Loop R1). In the absence of direct, reliable, and trusted measurements of customer satisfaction, management interprets the reduction in time per order as productivity gains due to learning, and reduces labor requirements (the False Learning Loop B3). The reduction in service capacity further increases work pressure on the service-delivery personnel, which in turn reduces the time per order, thus locking the system into a vicious cycle (the Death Spiral Loop R2). Despite initial equilibrium *and* stationary demand, the simulations consistently showed erosion of the service standard. In 500 simulations the erosion rate of desired time per order over 200 weeks was, on aver-

[6] The normalized standard deviations (σ/μ) of customer orders and the nonabsent service capacity were less than 4%.

Figure 4 Feedback Structure of Erosion of Service Standard

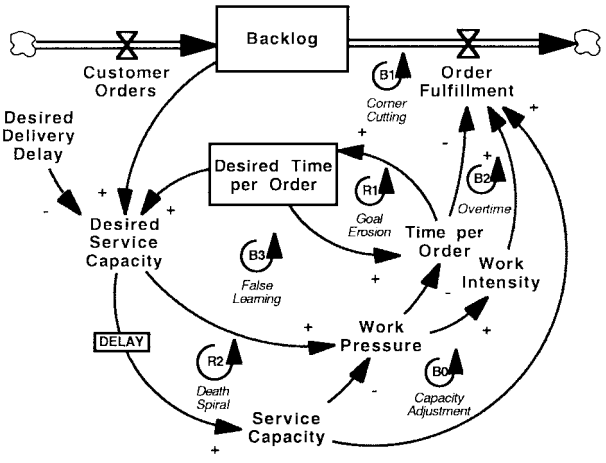
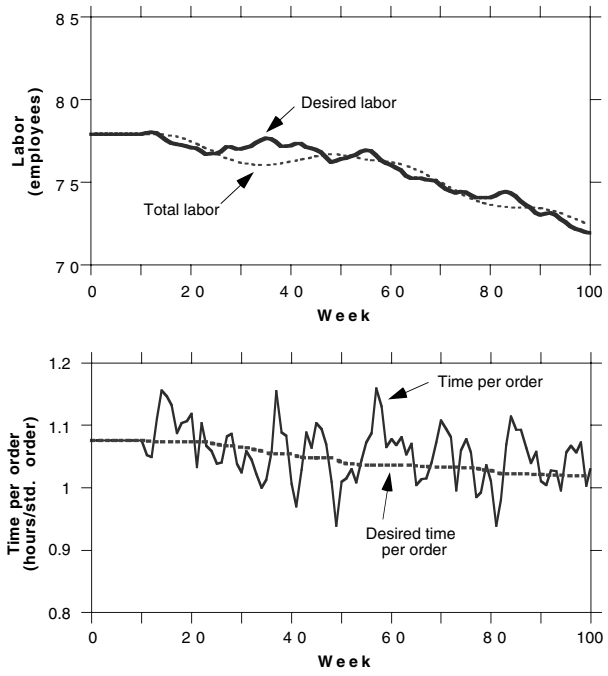


Figure 5 Response to Random Variations in Customer Orders and Absenteeism



age, 3.1% per year, a highly significant rate ($p \approx 0.00$).[7] Figure 5 shows the first hundred weeks of a typical simulation.

The observed erosion of the service standard could be explained by the lack of upward adjustment in desired time per order discussed above ($\tau_{ti} \approx \infty$). Though we found no evidence, either econometric or qualitative, for upward adjustment of the quality standard, it is nevertheless important to test the role of this assumption in the observed quality erosion. We found that even minor asymmetries in the standard formation process can lead to significant quality erosion. In simulations with a 10% difference between the upward and downward time constants for the adjustment of desired time per order ($\tau_{ti} = 1.1^{*}\tau_{td} =$ 20.7 weeks), the service standard still eroded at an average rate of 0.5% per year ($p \approx 0.03$). With fully symmetric adjustment ($\tau_{td} = \tau_{ti} = 18.8$ weeks) and stationary demand the erosion rate was 0.3% per year, but this value is not significantly different from zero ($p \approx 0.15$). However, this result is highly sensitive to the assumption of stationary demand. Simulating the system with modest demand growth of 3% per year, the target growth rate for Nelson House, caused quality to drop at an average rate of 1.7% per year ($p \approx 0.00$) even when quality norms adjust upward as readily as they adjust downward. Similarly, cutting normal employee tenure to 200 weeks, a value consistent with the drop in unemployment after the recession at the time of the study ended, causes average quality erosion 0.5% per year ($p \approx 0.04$), even without demand growth. The tendency toward quality erosion is not contingent on the assumption that quality norms decay readily, but rise only with difficulty.

[7] Throughout this section we report the average annualized erosion rates of the service standard after 210 weeks in a sample of 500 simulations; the p values report significance levels, under one-tailed tests, for H_0: erosion rate = 0.

4.2. Response to Work Pressure

To illustrate how the three responses to work pressure—increasing service capacity (SC), reducing time per order (TPO), and increasing work intensity (WI)—interact to generate the erosion of the service standard, the model was initialized in equilibrium and tested, without noise, with a 10% step increase in customer orders. Figure 6 shows the contribution to throughput from each of the responses, along with the change in throughput resulting from service standard erosion. The combination of responses is effective in immediately increasing throughput by 10%. However, the timing and strength of these responses differ substantially.

First, the initial reduction of TPO (Loop B1 in Figure 4) is almost twice as aggressive as the increase in WI (Loop B2). We found that workers under pressure to increase output are much more willing to cut corners (reduce the time they devote to

Figure 6 Response to a 10% Increase in Demand

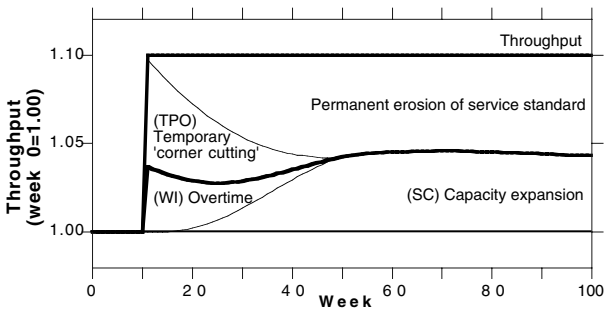


Table 4 Responses to Work Pressure and Consequences

	Response		Consequence		
	Elasticity	Average Response Lag[(1)] (Weeks)	Affected State Variable	Average Response Lag (Weeks)	Average Perception Lag (Weeks)
TPO response	−0.64	0	Desired time per order	$\tau_{td} = 18.7$	∞[(2)]
WI response	0.37	0	Fatigue	$\tau_{fe} = 3.0$	$\tau_{pe} = 6.7$
SC response	0.16	18.8 + 11.5 + 29.9	Service capacity	0	0

(1) Adjustments of TPO and WI are instantaneous once work pressure is identified. The SC response includes three successive delays: the time to adjust desired labor (τ_l^*), time to adjust labor (τ_l), and the hiring delay (τ_h).

(2) The effects of desired time per order are not detected in the LC because of the lack of quality metrics.

each customer), and only reluctantly work longer hours. Although in interviews and surveys employees claimed a deep concern for the "standard of customer service," no operational metrics of service quality were in place during the time of the study. Of the 15 loan officers interviewed, all but one admitted to reducing their effort to document transactions and to sell additional products in times of high work pressure. The weak response of quality pressure and the resulting willingness to cut time per order are consistent with the emphasis the monitoring system places on processing customer orders the same day they arrive.

Second, whereas employees' responses to work pressure—corner cutting and overtime—are essentially instantaneous, the adjustment of service capacity (Loop B0) is slow, peaking after 25 weeks. There are several reasons for the lag. First, although performance metrics are available on a weekly basis, they are summarized and analyzed at the end of the month; management must then decide how to update their estimate of labor productivity in assessing capacity requirements. Consistent with these practices, the statistical estimates showed that management perceives and reacts to changes in labor productivity with an average lag of $\tau_{pe} = 6.7$ weeks. Next, to smooth out the high-frequency variations in customer orders, management adjusts their estimate of required service capacity with an average lag of 4 months ($\tau_{l^*} =$ 18.8 weeks). The delay in adjusting authorized labor achieves its purpose of filtering out variations in customer orders (see desired labor and orders processed in Figure 3), and is consistent with management's imperative to control costs. Once labor is authorized it takes, on average, 7 months for the hiring process to bring a new employee into the LC ($\tau_h = 29.9$ weeks). Finally, we found rookies to be only 35% as productive as experienced personnel, with an average delay of about one quarter to become productive ($\tau_e = 12$ weeks). The combination of cautious hiring policies, hiring delays, and long training requirements mean service capacity is slow to react to changes in demand. Temporary variations in work pressure must therefore be accommodated by overtime or quality erosion.

The relative strength and timing of the responses (TPO > WI > SC) explains 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 capacity has eroded with the new service standard, and the model reaches equilibrium at a permanently lower quality level. In this particular test, the simulated organization increased its throughput 10% by reducing the internal standard of customer service 5.4% and increasing service capacity 4.1%.

The elasticity and lags discussed above are summarized in Table 4. The right side lists the state variables affected by each response, the time constant for the effect to take place, and the time it takes management to perceive those changes. Comparing the time constants for the consequences of each response, it becomes clear why TPO and WI are the preferred reactions: They proved instantaneous flexibility without any *apparent* cost. A change in service capacity, on the other hand, takes time (justifying, authorizing, hiring, and training new workers), but increases costs immediately. The preference for TPO over WI becomes clear when comparing the time it takes each

Table 5 Effect of Time per Order on Sales[†]

SQRT(Business Loan Volume)	$= a_0 + a_1$	TPO
	$= -719 +$	778 TPO
SE	(304.7)	(287.1)

$n = 50$; 10 left-censored observations (volume ≤ 0); $\chi_1^2 = 7.02$ ($p < .008$)

[†]Using TOBIT estimation. Results are also significant ($p < 0.01$) under ordinary least squares.

response to have a long-term effect on the performance of the lending center and the time it takes management to perceive it. Management can detect and respond to changes in productivity, but the lack of metrics for service quality prevents them from realizing the costs of eroding the service standard. The slow adjustment of capacity means random increases in demand can lead to cuts in service, cuts that gradually become embedded in employee norms for service. The longer the delay in adjusting capacity or the more flexible the norm for service, the larger the proportion of the demand increase absorbed by service erosion rather than capacity expansion.

4.3. Consequences

Does the erosion of service quality matter? Because customer service expectations adjust to past performance, it could be argued that a reduction in service standards represents productivity gains and is an effective cost-reduction strategy. The downward adjustment of service quality, however, implies a transitional dissatisfaction; customers will become habituated to lower expectations only after having experienced what they consider poor service. The long time constants associated with the adjustment of expectations suggest extended periods of time during which customers would be dissatisfied, predisposing them to consider alternative service providers.

There are, in addition, some immediate and tangible implications of reducing the service standard. Table 5 shows the estimated effect of the time spent with each customer on sales of business loans by the LC. Despite the large variance in the sales data, time per order is a significant predictor of loan volume (measured in £/week). The 4.1% reduction of the service standard during the period for which data were available implies a 50% reduction in expected sales. Lost sales, as large as they are, underestimate the hidden costs of a low service standard, as high work pressure also translates into errors in documentation and higher rework rates.

5. Policy Analysis

In this section, we explore policies to maintain service quality without compromising the organization's ability to respond to demand fluctuations. Parameters are as above, with the following exceptions. First, during the study the U.K. economy was in recession, suppressing employee turnover. Labor mobility increased when the labor market tightened. We therefore reduce average employee tenure to 4 years ($\tau_a^* = 200$ weeks). Second, though we found no evidence for any upward flexibility of quality norms, we allow desired time per order to increase by setting $\tau_{ti} = 1.5^* \tau_{td}$ (see discussion and analysis in §4.1). All simulations were run from initial equilibrium with random variations in demand and absenteeism introduced as specified in §4.1. These base-case assumptions generate average service-quality erosion of 1.28%/year (see Table 6), for the same reasons discussed in §4.1. A documented version of the model is available for experimentation under different assumptions.[8]

Expediting the Adjustment of Capacity. 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, reducing the delays governing the Capacity Acquisition Loop B0 in Figure 4. To test this policy, we reduced the time to adjust labor and the hiring delay by 50%, representing significant reengineering of the labor supply chain. The policy has a limited impact, reducing the quality erosion rate to l.07%/year, 16% less than the base case. Note, however, that this erosion rate is not significantly different from the base-case rate (Policy 1, Table 6).

Another strategy to increase the responsiveness of service capacity is to hire employees with greater

[8] http://www.people.hbs.edu/roliva/research/service/esq.html.

Table 6 Policy Analysis—Desired Time per Order Erosion Rate

Policy	Parameter Changes	Quality Erosion[a] (%/year)	p value H_0: ($e_r = 0$)	p value H_0: (e_r = Base Case)	Average Delivery Delay (weeks)
Base case		−1.28	0.000		0.100
(1) Faster capacity acquisition	$\tau_l = 6, \tau_h = 15$	−1.07	0.000	0.260	0.100
(2) Faster learning	$\tau_e = 6, \varepsilon = 0.6$	−1.33	0.000	0.445	0.100
(3) Reduced effect of work pressure	$f_{wt} = e^{-0.37w}, f_{wi} = e^{0.63w}$	−0.93	0.000	0.115	0.100
(4) Quality pressure (QP)	$f_{pt} = e^{1.0p}$	−1.05	0.000	0.229	0.101
(5) QP + management pressure	(4) & $\omega_e = 0.5, Q_m^* = 1$	−0.77	0.000	0.052	0.101
(6) QP + upward management pressure	(4) & $\omega_e = 0.5, Q_m^* = 1.05$	0.86	0.000	0.000	0.102
(7) Combined policy	(1) & (3) & (6)	1.39	0.000	0.000	0.101

(a) The reported rates are the average annualized erosion rate of the employee's quality standard (desired time per customer) after 210 simulated weeks over 500 simulations. The p values report significance levels, under one-tailed tests, for H_0: erosion rate = 0 and, under two-sample normal model with unequal variances, for H_0: erosion rate = erosion rate of the base case.

initial effectiveness or to accelerate their learning process. Unfortunately, these options are rarely available in high-contact services that require job-specific knowledge. Nevertheless, to test the potential for this policy, we optimistically assume initial effectiveness rises from 35% to 60% of the productivity of experienced workers and that the learning period is cut by 50%. Despite these large changes, this strategy has a negligible impact, leaving the erosion rate essentially unchanged (Policy 2, Table 6). The policy has low leverage because the assumption of stationary demand implies the steady-state rookie fraction is quite low (about 6% of the workforce). Policies that speed the learning curve will, however, be more effective in start-up conditions or high-growth industries, when large numbers of rookies can overwhelm a service organization.

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 customer caused by high work pressure. We found that employees at our site were twice as willing to cut corners as to work overtime. Reducing their willingness to cut corners should weaken the Goal Erosion Loop and slow the decline of the service standard. 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 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.[9] This could 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). We assume such process changes and incentives cause workers to be twice as willing to use overtime as to cut corners ($f_{wt} = e^{-0.37w}$ and $f_{wi} = e^{0.63w}$). The average erosion rate falls by 27% to −0.93%/year (Policy 3, Table 6). Quality erodes even when overtime is the priority because high work pressure still causes employees to cut the time they devote to each customer; these temporary reductions then gradually drag down the norm for time per order. The Goal Erosion Loop R1 is weaker, but still unopposed.

Creating Quality Pressure. Our fieldwork revealed that there was no effective pressure from quality norms to counteract cuts in service induced by high work pressure, even after work pressure returned to normal. Though loan officers reported

[9] Since the overall response to work pressure is given by $c(e^{\beta w}/e^{\alpha w})$ (substituting Equations (22) and (25) in Equation (2)), $\beta - \alpha = 1$ ensures full responsiveness to work pressure.

some discomfort with their performance, we found no evidence that low quality had any impact on the time employees devoted to each customer (technically, the estimated elasticity of time per customer with respect to quality pressure was zero; see Table 2).

Creating quality pressure requires management to become aware of the implications of poor service—lost sales, rework, and customer defections—and then 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. We simulate such programs by assuming workers boost the time allocated to each customer whenever the quality they perceive falls below their standards. We optimistically assume a response to quality pressure ($f_{pt} = e^{1.00p}$), equal to the combined responses to work pressure ($f_{wt} = e^{-0.64w}$ and $f_{wi} = e^{0.37w}$). This policy creates a new balancing feedback loop that attempts to eliminate gaps between the standard and perceived time per customer by boosting the actual time spent with each customer request.

However, the policy has only a small effect, reducing the quality-erosion rate by 18%, to −1.05%/year, a value not significantly different from the base case (Policy 4, Table 6). The policy fails for three reasons. First, it is fundamentally reactive: Quality pressure works to increase time per customer only *after* high work pressure has forced workers to cut the time they spend on each customer below standards. Second, to the extent workers do respond to low quality and increase the time allocated to each customer, throughput falls. As work pressure builds, employees are forced to spend less time with each customer to clear the backlog. Note that the policy increases delivery delay by an average of 1%, with delays rising by as much as 10% during peak periods. Finally, the policy does not halt the erosion of the workers' standards for service. Despite the aggressive response to quality pressure, time per customer still drops when work pressure rises, gradually dragging employee standards down, and therefore dissipating quality pressure.

Quality erosion is not avoided even when employees are highly responsive to any drop in quality relative to their standards. It is also necessary to prevent the erosion of their standards. An external norm for service quality may provide sufficient counterpressure to halt standard erosion. In some industries such external norms may be developed as part of the professional training of service providers (health care provides a—perhaps debatable—example). More often, management must take an active role in the formation of the service standard by articulating clear and consistent standards for service quality unaffected by the organization's own past performance, and then monitoring performance against them.

We simulate a focus on external norms by altering the employees' standard formation process to include the influence of management's quality goal Q_m^*. We set management's quality goal to one, representing the quality level that satisfies customer needs. This value might correspond to an aspiration of "zero defects" (no complaints). How much weight should the external norm receive relative to the employees' own experience? Because the service encounter is essentially personal, intangible, and negotiated between server and customer, it cannot be fully standardized. Employees' experience will continue to form an important input into their beliefs about how and how much time they should spend with each customer. To test the policy we assume that the weight accorded to management's quality goal rises from zero to 50% ($\omega_e = 0.5$; see Equation (33)).

The addition of an external reference for quality goals further slows the quality-erosion rate, which falls to an average of −0.77%/year, a drop of 40% from the base case (Policy 5, Table 6). Yet, the policy is not able to stop quality erosion altogether. While the external quality goal weakens the reinforcing Goal Erosion Loop (R1), the impact of quality pressure is still fundamentally reactive: It offsets the impact of work pressure only when perceived quality drops below the standard. A policy of aggressive quality pressure, even with an external goal of full customer satisfaction, cannot have any impact until at least some customers are dissatisfied.

To arrest quality erosion before customers are dissatisfied, management must strive to exceed customer expectations. Policy 6 in Table 6 tests this policy of "stretch objectives" by repeating Policy 5 while setting management's quality goal above one ($Q_m^* = 1.05$, representing the goal of delighting the customers, not

merely satisfying them). This policy results in a rise in quality of 0.86%/year, a highly significant result. However, as expected, the policy increases the time employees spend on each customer and forces delivery delays up (by an average of 2%). Slow service itself can degrade customers' experience and cause them to defect. In addition, the buildup of work pressure still counteracts the benefits of quality pressure.

Policy 7 addresses the delivery-time issue by combining Policy 6 with faster capacity acquisition (Policy 1) and the reduced effect of work pressure on time per customer (Policy 3). Faster capacity acquisition should reduce the duration of any periods during which work intensity is high; reducing the effect of work pressure on time per customer further weakens the goal-erosion process and augments effective capacity by boosting employees' willingness to work overtime during peak periods. The combination policy results in quality improvement of about 1.4%/year and reduces the average delivery delay compared to Policy 6. Note that the combined impact is less than the sum of the individual impacts: Diminishing returns result from the strong compensating negative feedbacks controlling work pressure and quality.

6. Implications

Despite the quality revolution of the past two decades, the quality of service in many industries has eroded. To understand how service quality could persistently erode, we developed a dynamic model of a service organization. The model provides an endogenous account of service delivery that integrates physical, institutional, economic, and psychological factors to explain how service throughput and quality evolve as demand and capacity vary. We used a wide range of data from the field study, including data on order flows, service capacity, management hiring practices, and overtime to estimate the strength of the hypothesized relationships and the behavioral responses of managers, employees, and customers. The model was tested by statistically comparing its behavior against multiple historical data series.

The theory builds on organizational learning models in the tradition of Cyert and March (1963), Levinthal and March (1981), and others. The agents in the model (workers, managers, and customers) are portrayed as boundedly rational (Morecroft 1985, Simon 1957), but also as social beings who respond to the norms and behaviors of those around them. The decision rules of the agents are grounded in well-established research in the behavioral decision-making, organizational-learning, and system-dynamics literature, including the aspiration-adjustment process (Lant 1992), anchoring and adjustment (Hogarth 1980), and hill climbing as a learning process. Our work supports studies showing that learning can lead to dysfunctional outcomes and threaten organizational survival (Forrester 1961, March 1991, Masuch 1985, Sastry 1997, Sterman et al. 1997). The theory differs from some prior models in integrating these heuristics with a dynamic, disequilibrium account of the physical and institutional structure of the organization, including hiring delays, on-the-job training and mentoring, workflow, and task backlog. The interaction of the actors with one another and with the disequilibrium pressures in their physical and institutional environment leads to unintended and dysfunctional dynamics. We further show how the learning processes of the agents lead them to intensify the disequilibrium pressures, trapping the organization in a vicious cycle of declining quality. Our work moves beyond most existing studies by tightly grounding our assumptions about decision-making processes in a detailed field study. Finally, we use the grounded and calibrated model to develop and test policy recommendations aimed at avoiding or reversing these dysfunctional outcomes.

The form of dysfunctional learning we identify—service-quality erosion—has increasing managerial and economic significance as the share of the global economy consisting of services grows and as evidence of service-quality erosion accumulates. We found that service quality can erode, even under stationary demand, due to a reinforcing feedback that arises from intendedly rational decisions by each actor in a service setting. Employees, in an effort to meet throughput goals, absorb small variations in workload by reducing the time spent with each customer and by working longer hours. The reduction in time per customer, while enabling an immediate increase in throughput, gradually erodes service norms in the organization. In the absence of direct and reliable

measurements of customer satisfaction, and consistent with their imperative to control cost, management interprets the reduction in time per order as a productivity gain and reduces the labor force. The drop in service capacity further increases the workload, so service personnel are forced to cut the time per customer still more. These factors interact to generate the potential for significant, ongoing quality erosion, even when resources are on average sufficient to meet demand. The consequences of such erosion are potentially severe: Besides the obvious costs of corner cutting (poor documentation, rework, customer defection, etc.), we found that inadvertent cuts in the time loan officers spent working with customers led to large and statistically significant drops in sales of ancillary services. The results were lower profit, slower growth, and greater financial pressure on the organization to boost productivity, further intensifying the workload and the pressure to cut corners.

An alternative explanation for eroding service quality is increasing customer expectations—perhaps resulting from exposure to better levels of service in other industries. A fortiori we assumed constant customer expectations, thus generating an endogenous explanation for erosion of service quality. The erosion of service standards in high-contact services is the result of the relative intensity of the available responses to work pressure and the absence of a fixed objective standard. The relative intensity of the responses is determined by the structural characteristics of high-contact services, specifically the need to customize service transactions and the delays in developing employee skills. Customization inhibits the standardization of the service-delivery process, 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 training delays and learning curves, delays in capacity expansion, and the intangibility of service quality, the structure that can lead to quality erosion is likely to be common throughout the service sector.

While our field study centered on a labor-intensive setting, the theory and the tendency toward erosion of quality standards are not limited to high-contact services. For example, online trading and other Internet businesses have been facing unexpectedly high rates of demand growth. Many believe standardized and automated e-commerce transactions offer a consistent high-quality service interaction for all. Yet many e-businesses faced with high levels of work pressure find themselves unable to provide adequate support, that is, customize the service interaction, when something goes wrong or as customer needs evolve. The consequences include higher cost, loss of reputation and market share, and slower growth, all affecting market valuation or even survival. Beyond the application of this framework in other settings, future research should strive for theoretical enrichment, expanding the model boundary to include financial pressures, market dynamics, and other dimensions of service quality. Although not relevant for the bank setting, further exploration of the responses to work pressure should include customer responses to low quality or delays in service (e.g., balking) and dynamic pricing mechanisms (e.g., yield management) to regulate demand.

Acknowledgments

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.

References

American customer satisfaction index. 2001. American Society for Quality, ⟨http://acsi.asq.org/⟩.

Argote, L., D. Epple. 1990. Learning curves in manufacturing. *Science* **247** 920–924.

Barlas, Y. 1989. Multiple tests for validations of system dynamics type of simulation models. *Eur. J. Oper. Res.* **42**(1) 59–87.

Barnett, W., M. Hansen. 1996. The red queen in organizational evolution. *Strategic Management J.* **17** 139–157.

Baumol, W. 1967. Macroeconomics of unbalanced growth: The anatomy of urban crisis. *Amer. Econom. Rev.* **57**(June) 415–426.

———, S. Blackman, E. Wolf. 1991. *Productivity and American Leadership*. MIT Press, Cambridge, MA.

Boulding, W., A. Karla, R. Staelin, V. Zeithaml. 1993. A dynamic process model of service quality. *J. Marketing Res.* **30**(1) 7–27.

Broh, R. 1982. *Managing Quality for Higher Profits*. McGraw-Hill, New York.

Chase, R. 1981. The customer contact approach to services: Theoretical bases and practical extensions. *Oper. Res.* **29**(4) 698–706.

Cyert, R., J. March. 1963. *A Behavioral Theory of the Firm*. Prentice Hall, Englewood Cliffs, NJ.

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

Einhorn, H., R. Hogarth. 1981. Behavioral decision theory: Process of judgment and choice. *Annu. Rev. Psych.* **32** 53–88.

Farber, B., ed. 1983. *Stress and Burnout in the Human Service Professions*. Pergamon Press, New York.

Forrester, J. 1961. *Industrial Dynamics*. MIT Press, Cambridge, MA.

———. 1979. An alternative approach to economic policy: Macrobehavior from microstructure. N.M. Kamrany, R.H. Day, eds. *Economic Issues of the Eighties*. The Johns Hopkins University Press, Baltimore, MD, 80–108.

———, P. Senge. 1980. Tests for building confidence in system dynamics models. *TIMS Stud. Management Sci.* **14** 209–228.

Hogarth, R. 1980. *Judgment and Choice: The Psychology of Decision*. Wiley, New York.

Homer, J. 1985. Worker burnout: A dynamic model with implications for prevention and control. *System Dynam. Rev.* **1**(1) 42–62.

Jarmain, W., ed. 1963. *Problems in Industrial Dynamics*. MIT Press, Cambridge, MA.

Kahneman, D., A. Tversky. 1982. The psychology of preferences. *Sci. Amer.* **246** 160–173.

Koepp, S. 1987. Why is service so bad? Pul-eeze! Will somebody help me? *Time* (Feb. 2) 46.

Lant, T. 1992. Aspiration level adaptation: An empirical exploration. *Management Sci.* **38**(5) 623–644.

Levinthal, D., J. March. 1981. A model of adaptive organizational search. *J. Econom. Behavior and Organ.* **2**(4) 307–333.

March, J. 1991. Exploration and exploitation in organizational learning. *Organ. Sci.* **2**(1) 71–87.

Masuch, M. 1985. Vicious cycles in organizations. *Admin. Sci. Quart.* **30** 14–33.

McKinsey Global Institute. 1992. *Service Sector Productivity*. McKinsey and Company, Washington, DC.

Mills, P. 1986. *Managing Service Industries*. Ballinger, Cambridge, MA.

Mobley, W. H. 1982. *Employee Turnover: Causes, Consequences and Control*. Addison-Wesley, Reading, MA.

Morecroft, J. 1985. Rationality in the analysis of behavioral simulation models. *Management Sci.* **31**(7) 900–916.

Naylor, T., J. Finger. 1967. Verification of computer simulation models. *Management Sci.* **14**(2) 92–101.

Oliva, R. 1996. A dynamic theory of service delivery: Implications for managing service quality. PhD Thesis, Sloan School of Management, MIT, Cambridge, MA.

Sastry, M. 1997. Problems and paradoxes in a model of punctuated organizational change. *Admin. Sci. Quart.* **42**(2) 237–275.

Schneider, B. 1991. Service quality and profits: Can you have your cake and eat it, too?, *Human Res. Planning* **14**(2) 151–157.

———, J. Parkington, V. Buxton. 1980. Employee and customer perceptions of service in banks. *Admin. Sci. Quart.* **25**(2) 252–267.

Senge, P. 1990. Catalyzing systems thinking within organizations. F. Masaryk, ed. *Advances in Organizational Development*. Ablex, Norwood, NJ. 197–246.

———, J. Sterman. 1992. Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. *Euro. J. Oper. Res.* **59**(1) 137–150.

Shiba, S., A. K. Graham, D. Walden. 1993. *A New American TQM: Four Practical Revolutions in Management*. Productivity Press, Cambridge, MA.

Simon, H. A. 1957. *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in a Social Setting*. Wiley, New York.

Sterman, J. 1984. Appropriate summary statistics for evaluating the historical fit of system dynamics models. *Dynamica* **10** (Winter) 51–66.

———. 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Sci.* **35**(3) 321–339.

———, N. Repenning, F. Kofman. 1997. Unanticipated side effects of successful quality programs: Exploring a paradox of organizational improvement. *Management Sci.* **43**(4) 503–521.

Strandvik, T. 1994. Tolerance Zones in Perceived Service Quality. PhD Thesis, Swedish School of Economics and Business Administration, Helsinki, Finland.

Theil, H. 1966. *Applied Economic Forecasting*. North-Holland, New York.

Thomas, H. 1993. Effects of scheduled overtime on labor productivity: A literature review and analysis. Source Document, Pennsylvania State University, University Park, PA.

Tornow, W. W., J. W. Wiley. 1991. Service quality and management practices: A look at employee attitudes, customer satisfaction and bottom-line consequences. *Human Res. Planning*. **14**(2) 105–116.

U.S. Products get better markets than service. 1998. *Quality* **37**(1) 18.

van Horn, R. 1971. Validation of simulation results. *Management Sci.* **17**(5) 247–258.

Weisberg, J. 1994. Measuring worker's burnout and intention to leave. *Quality of Working Life.* **15**(1) 4–14.

Zangwill, W., P. Kantor. 1998. Toward a theory of continuous improvement and the learning curve. *Management Sci.* **44**(7) 910–920.

Zeithaml, V., L. Berry, A. Parasuraman. 1993. The nature and determinants of customer expectations of service. *Acad. Marketing Sci.* **21**(1) 1–13.

———, A. Parasuraman, L. Berry. 1990. *Delivering Quality Service: Balancing Customer Perceptions and Expectations*. Free Press, New York.

Accepted by Linda Argote; received July 19, 1999. This paper was with the authors 5 months for 2 revisions.