

**Cutting Corners and working overtime:
Quality erosion in the service industry**

Rogelio Oliva, Harvard Business School, and John
D. Sterman, MIT Sloan School of Management

#4075-99-MSA

June, 1999

Cutting corners and working overtime: Quality erosion in the service industry

Rogelio Oliva
Harvard Business School
Baker Library 182
Boston, MA 02163
Phone 617-495-5049; Fax 617-496-7167
roliva@hbs.edu

John D. Sterman
MIT, Sloan School of Management
30 Wadsworth St, E53-351
Cambridge, MA 02142
Phone 617-253-1951; Fax 617-258-7579
jsterman@mit.edu

DRAFT FOR COMMENT ONLY – NOT FOR QUOTATION OR CITATION

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

Abstract

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 1997 to 67.7, down nine percentage points from its 1994 value. We hypothesize that the characteristics of services – inseparability, intangibility, and labor intensity – interact with management practices to bias service providers to reduce the level of service they deliver, often locking entire industries into a vicious cycle of eroding service standards. To explore this proposition we developed a formal model that integrates the structural elements of service delivery. We used econometric estimation, interviews, observations, and archival data to calibrate the model for a consumer lending service center in a major UK bank. 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.

(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 drift toward 'high-contact' services (Chase 1981) – the challenges that service managers are facing 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. Since there is no agreed objective standard about the service to be delivered, the only criteria available to evaluate service quality is a subjective comparison of customers' expectations to their perception of the actual service delivered (Gummesson 1993; Zeithaml, Parasuraman et al. 1990). Furthermore, customers do not evaluate service quality solely in terms of the outcome of the interaction; they also consider the process of service delivery. Hence, service quality is a multidimensional construct that encompasses all aspects of service delivery, and is difficult to assess and communicate.

The lack of clarity about customers' requirements also translates into operational difficulties. Service capacity, in a setting where most of the value is added by the interpersonal exchange between consumer and provider, is a function of the provider's skills, efforts and attitudes. Because of the intangibility of services, it is not always possible to articulate the set of skills and attitudes that customers value. Consequently, it is difficult to hire the right set of skills, and most service providers develop them through lengthy on-the-job training.

Second, services are typically produced in the presence of the customer and customers often participate in the production process. The instantaneous nature of the provision of service and its consumption brings 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 of service quality. Studies have shown a positive relationship

between the perceptions, attitudes and intentions of employees and customers (Schneider and Bowen 1985; Schneider, Parkington 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 co-evolution of their perceptions and expectations. Evolving customer expectations suggest the need for a dynamic process to adjust service specifications to match their ever-changing needs (Boulding, Karla et al. 1993).

Finally, the high degree of customization and value created through the personal interaction of customers and service providers mean significant productivity gains cannot be expected through capital substitution in high-contact services. Baumol (1967; Baumol, Batey Blackman 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 – the cost disease hypothesis. Increasing unit cost translates into financial pressure on the firms of 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 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. Since 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 to minimize 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. As entire industries become locked in cycles of underinvestment and eroding standards, industry norms reinforcing expense control and productivity become increasingly influential in shaping individual firm decisions. Industry-wide erosion of service quality has been frequently cited in the popular press (e.g. Quality 1998b; Koepp 1987), and recently reported by the American Customer Satisfaction Index. The 1997 ACSI index for services fell to 67.7, down nine percentage points from its 1994 value (American Society for Quality 1998a).

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 our research approach. First, we developed a formal model that integrates the structural elements of service settings (§2). We then tested the model empirically through calibration to a research site – a consumer lending service center in a major UK bank (§3 and 4). We then used the model to understand the sources of service quality erosion and generate some policy recommendations (§5). Finally, we discuss the implications of our findings for the service industry in general and identify future research areas.

2. Model Structure

In this section we present a model of a service center that allows us to test the hypotheses described in the previous section. Theoretical foundations and evidence for the hypothesized causal relationships are presented with each equation. The evidence is mainly drawn from the human resources, behavioral decision theory, marketing, operations management, and system dynamics literature. The purpose of the model is to test whether the structural elements of a service delivery

process – physical flows, organizational structure and decision making – are responsible for the tendency towards quality erosion identified in the service industry (see figure 1 for overall model structure). The focus of this inquiry is operational; the goal is to provide an explanation and identify policies that would permit management to intervene and improve performance.

Service delivery. Customer orders (s_o) are not processed immediately and accumulate in a backlog (B) before being processed. The order rate is exogenous. Exogenous orders imply 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).

$$(1) \quad (d/dt)B = s_o - s_f$$

The order fulfillment rate (s_f) is the effective service capacity (c) corrected 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). Service capacity is measured in homogeneous 'capacity hours' required to fulfill each customer order at a given quality level. 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).

$$(2) \quad s_f = \min(c \cdot i/T, B/\tau_f)$$

A single factor of production (labor) is considered.¹ Although the units of service capacity are assumed to be homogeneous, 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 the workers' skills and efforts. Effective service capacity (c) is determined by adjusting the total labor force (L) by the effects of personnel experience (e) and fatigue (f).

$$(3) \quad c = L \cdot e \cdot f$$

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

$$(4) \quad c^* = (B/\lambda) \cdot T^*$$

Work pressure (w), a metric of the relative workload in the service center, is defined as the gap between required service capacity and effective service capacity as a fraction of current capacity.

$$(5) \quad w = (c^* - c)/c$$

Since consumption and delivery of services are simultaneous, it is difficult to strike a balance between supply and demand. Whereas excess service capacity incurs higher operating costs, shortfalls in service capacity are absorbed by the backlog of unfilled customer orders, generating higher work pressure. There are only three ways to increase the order fulfillment rate (Senge and Sterman 1992): 1) increase work intensity, 2) reduce the time allocated to processing each customer order, or 3) expand service capacity. The first two responses can be implemented directly by employees performing the service. Expanding service capacity, however, is normally a managerial decision and takes time.

Responses to work pressure. The first response to an increase in work pressure is for employees to increase their work intensity, by taking shorter breaks and working overtime. Work intensity is greater than one in the case of overtime. In the model it is assumed that employees adjust work intensity (i) in response to work pressure (w). The response is non-linear and limited by the amount of time an employee could be working.

$$(6) \quad i = f_{wi}(w) \quad f(0) = 1, f(\infty) = i^{\max}, 0 \leq f \leq 1$$

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 non-linear function that reduces effective service capacity when service personnel are tired (see eq. 3).

$$(7) \quad (d/dt)F_e = (i - F_e)/\tau_{fe}$$

$$(8) \quad f = f_{fe}(F_e) \quad f(F_e \leq 1) = 1, f \leq 0, f'' > 0$$

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 effects of fatigue on employee attrition (a_f). The time constant for the fatigue level during attrition is τ_{fa} . While extended overtime soon affects productivity, the impact of burnout on attrition is slower. Hence, $\tau_{fa} > \tau_{fe}$.

$$(9) \quad (d/dt)F_a = (i - F_a)/\tau_{fa}$$

$$(10) \quad a_f = f_{fa}(F_a) \quad f(F_a \leq x) = 1, f(\infty) = 0, f' \leq 0$$

The second way to deal with high work pressure is cutting the time allocated to each order (T). An anchoring and adjustment process (Einhorn and Hogarth 1981) is assumed. Employees decide on a service level by anchoring on a pre-existing service standard (the anchor), and adjusting it to take account of current workload (t_w) and quality pressure (t_p). In turn, the current performance level 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_p).

$$(11) \quad T = \max(t_w \cdot t_p \cdot T^*, \tau_p)$$

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 non-linear and to be neutral in the absence of pressure.

$$(12) \quad t_w = f_{wt}(w) \quad f(0) = 1, f' \leq 0$$

$$(13) \quad t_p = f_{pt}(p) \quad f(0) = 1, f' \geq 0$$

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 asymmetrical. Asymmetrical 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.

$$(14) \quad (d/dt)T^* = (T - T^*)/\tau_{to} \quad \tau_{to} = \begin{cases} \tau_{ti} & \text{If } T > T^* \\ \tau_{td} & \text{otherwise;} \end{cases}$$

Expanding service capacity to meet demand is the third option for reducing work pressure. 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.

$$(15) \quad (d/dt)E = ((c/L) - E)/\tau_{pe}$$

$$(16) \quad (d/dt)L^* = ((c^*/E) - L^*)/\tau_l$$

Service capacity. Learning by doing is well documented in a wide range of settings, including service delivery organizations (Argote and Epple 1990; Darr, Argote 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. Because these services involve personal and customized interaction between individual servers and customers we expect 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 that recently hired personnel undergo (developing their skills) as an “experience chain” (Jarman 1963). Recently hired personnel are assumed to have only a fraction (ϵ) of the productivity of more experienced employees, but through on-the-job coaching, mentoring, and experience, 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) utilized in the derivation of service capacity (see eq. 3). The effect of experience is the number of full-time equivalent experienced personnel relative to the total labor force.²

(17)

$$L = L_e + L_r$$

(18)

$$e = \max\left(0, (L_e + L_r(\varepsilon - \eta))/L\right) \quad 0 \leq \varepsilon \leq 1, \eta \geq 0$$

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

(19)

$$(d/dt)L_r = l_h - l_e$$

(20)

$$(d/dt)L_e = l_e - l_a$$

(21)

$$l_e = L_r/\tau_e$$

The turnover from the experienced personnel stock is assumed exponential with an average time for turnover (τ_a). The training period is relatively short in comparison to the average tenure of employment; hence we ignore turnover from the rookie stock. Two factors modify the average time for turnover: employees' fatigue (a_f) (see eq. 10) and perception of service quality (a_q) (see eq. 35). Although research has identified determinants of labor turnover depending on the external economy, organizational attributes, and individual factors (Mobley 1982), most of these factors are shared by all the employees in a service operation and thus considered exogenous to the model and captured in the nominal time for turnover (τ_a^*).

(22)

$$l_a = L_e/\tau_a$$

(23)

$$\tau_a = \tau_a^* \cdot a_f \cdot a_q$$

It takes time to hire new employees. The hiring rate depends on the firm's unfilled labor vacancies (L_v) and a hiring delay (τ_h). Vacancies represent the labor orders (l_o) that have not been filled. By

Little's law, desired vacancies (L_v^*) are assumed to be proportional to the desired hiring rate and the hiring delay (τ_h), the time it normally takes to fill a vacancy.

$$(24) \quad l_h = L_v / \tau_h$$

$$(25) \quad (d/dt)L_v = l_o - l_h$$

$$(26) \quad L_v^* = l_h^* \cdot \tau_h$$

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

$$(27) \quad l_o^* = l_h^* + (L_v^* - L_v) / \tau_l$$

$$(28) \quad l_h^* = l_r + (L^* - L) / \tau_l \quad l_r = \begin{cases} 0 & \text{If } L > L^* \\ l_a & \text{otherwise;} \end{cases}$$

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).⁴

$$(29) \quad l_o = \max(-L_v / \tau_v, l_o^*)$$

Perceptions and expectations of service quality. To explore how the service center responds to changes in service quality, we model how the different actors perceive service quality, how service expectations are created, and how these perceptions and expectations affect behavior. To address the issues of service *inseparability* and *intangibility*, we defined service quality as a function of the time allocated per customer and customers' expectations. Since 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 they perceive a lack of skills or poor attitude. As more effective time is allocated to process each order, employees are able to inquire and satisfy customer needs beyond transactional requirements. The assumption that time per order is the main driver of service quality is consistent with Mills' (1986) equation of service quality with server productivity and a commonly made 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, Parasuraman 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^*). Experienced satisfaction or quality (q) is a non-linear function of the performance gap – the normalized difference between the time allocated per order (T) and customers’ expectations (Zeithaml, Parasuraman et al. 1990).

$$(30) \quad q = f_q\left((T - T_c^*)/T_c^*\right) \quad f(0) = 1, \quad 0 \leq f\{\cdot\} \leq f^{\max}, f' \geq 0$$

Although the exact relationship between effective time per order and service quality might vary from setting to setting, some generic characteristics can be specified. The existence of a “tolerance zone” for service quality (Strandvik 1994; Zeithaml, Berry et al. 1993) suggests a function that is relatively flat around zero but grows progressively steeper as the performance gap rises. Kano’s differentiation of quality attributes between *must-be’s* and *delighters* (Shiba, Graham et al. 1993) indicates that there are diminishing returns to the perceived value of an attribute, suggesting a saturation effect as performance rise 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 level perceived by employees (Q_e), management (Q_m) and customers (Q_c) adjusts via first-order exponential smoothing of actual quality. The time constants for these perceptual processes are assumed different and ranked according to their immediacy to the delivery process and the frequency of exposure to it.

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

Employees have a relatively accurate and speedy perception of the delivery process. Managers, through direct supervision or observation, have more removed perception of the service delivery process. Finally, even though customers have direct perception of the delivery process as individuals, they are not exposed to it as frequently as employees and it takes time for information

on quality to disseminate among consumers through word of mouth or external ratings (e.g. J.D. Power, Consumer Reports). Consequently, the average perception of quality in the market has an even longer adjustment constant, and $\tau_{qe} \leq \tau_{qm} \leq \tau_{qc}$.

In addition to the perception process described above, each agent involved in the service delivery process – employees, managers, and customers – is assumed to have an internal expectation of the service level that ought to be delivered. Customers hold an expectation (T_e^*) of what constitutes an adequate time to process their orders. Management has an explicit goal (Q_m^*) that reflects its desired level of quality. Finally, employees hold an internal quality standard (Q_e^*) that represents the level of quality they would deliver in the absence of work pressure and under normal work intensity. These expectations are conceptualized as levels of aspiration (Lant 1992; Lewin, Dembo et al. 1944; 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). Since the assessment of service quality is based exclusively on the comparison between perceptions and expectations, the aspiration adjustment process is particularly appropriate in the creation of quality expectations (Boulding, Karla et al. 1993).

The employees' quality standard (Q_e^*) is assumed to adapt 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^*). Management's quality goal (Q_m^*) is anchored to management's own perception of delivered quality (Q_m) and adapts to close the gap identified from the feedback from customers' perception of service quality ($1-Q_e$) – if the appropriate market research instruments are in place. Finally, customer expectations are anchored to the service provided by other suppliers in the industry (μ), and adapt to the current service experienced (T). An argument similar to that for the time constants of the perception process for employees, managers, and customers can be made for the time constants for the adjustment of aspirations, so $\tau_{ee} \leq \tau_{em} \leq \tau_{ec}$.

$$(32) \quad (d/dt)Q_e^* = (\omega_e Q_e + (1 - \omega_e)Q_m^* - Q_e^*)/\tau_{ee} \quad 0 \leq \omega_e \leq 1$$

$$(33) \quad (d/dt)Q_m^* = (Q_m + (1 - \omega_m)Q_e^* - Q_m^*)/\tau_{em} \quad 0 \leq \omega_m \leq 1$$

$$(34) \quad (d/dt)T_c^* = (\omega_c \mu + (1 - \omega_c)T - T_c^*)/\tau_{ec} \quad 0 \leq \omega_c \leq 1$$

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, Parkington et al. 1980). In this model, when employees perceive quality is low, the average duration of employment falls (see eq. 23).

$$(35) \quad a_q = f_{qa}(Q_e) \quad f(0) = 0, f(1) = 1, f' \geq 0$$

Second, the gap between employees' perception of delivered service quality (Q_e) and their quality expectation (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 (eq. 12).

$$(36) \quad p = (Q_e^* - Q_e)/Q_e^*$$

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 (eq. 32).

3. Empirical Testing

There are two chief concerns in testing a complex dynamic theory: Does the microstructure of the model correspond to what is known about the real system? Can the macro-behavior of the service setting be explained from the structural components of the theory? To test and build confidence in the model, its ability to replicate the behavior of a wide range of service settings should be assessed. As a first step in this process we tested the model against a particular service setting – a retail banking operation in the United Kingdom. Calibration tests the model's ability to capture the characteristics of the research site and illustrates its potential relevance to managers. The process, however, has some limitations. The site may not be typical, and there is a risk that not all the hypothesized relationships are active during the period for which data are collected. Validation, in this context, refers to the iterative “process for building an acceptable level of confidence” (van Horn 1971, pp. 247) in a model and its implications.

3.1. The Research Site

National Westminster Bank, Plc. is the flagship of NatWest Group, one of the largest financial institutions in the UK. The bank offers consumer and commercial banking at more than 2,200 locations. In 1990 the UK Retail Banking Services 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 percent of the total account volume of UK 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 the operational metrics of the service center; (2) interviews with employees, their managers, and staff, inside and outside the LC; (3) twelve hours of direct observation; and (4) archival data such as policy and procedures 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 remainder of this section presents an example of model estimation for a critical behavioral decision – the employees' allocation of time per order – and the use of data from multiple sources to make sense of the statistical results.

3.2. Partial model estimation

In the proposed model we hypothesize 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) (eq. 11). 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 (eqs. 11 & 14), and the “goal adjustment” that occurs as desired time per order determines required service capacity (eqs. 4, 5, 11, 12 & 14). 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 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 (eq. 13'). The partial model 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):

$$\begin{aligned}
 & \underset{T_0^*, \alpha, \tau_{ii}, \tau_{id}}{\text{Min}} \sum_{t=1}^n (T(t) - TPO(t))^2 \\
 & \text{Subject to} \\
 & T(t) = \max(t_p(t) \cdot t_w(t) \cdot T^*(t), \tau_f); \quad \tau_f = 0.1 \quad (11') \\
 & T^*(t) = \int (T(t) - T^*(t)) / \tau_{to} + T_0^*; \quad \tau_{to} = \begin{cases} \tau_{ii} & \text{If } (T(t) > T^*(t)) \\ \tau_{id} & \text{otherwise;} \end{cases} \quad (14') \\
 & sc^*(t) = CO(t) \cdot T^*(t) \quad (4') \\
 & w(t) = (sc^*(t) - SC(t)) / SC(t) \quad (5') \\
 & t_w(t) = \exp(\alpha w(t)) \quad (12') \\
 & t_p(t) = 1 \quad (13')
 \end{aligned}$$

We derived the service capacity (SC) data series from the number of employees corrected for absenteeism and adjusted for the effects of fatigue and experience.⁵ Since the LC cleared the backlog of orders every day, customer orders (CO) served as a proxy for the desired fulfillment

rate (B/λ ; eq. 4'). 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 percent 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. 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 (Sternan 1984).

The initial estimate for desired time per order is 1.08 person-hours, about 8% less than management's goal (bank procedures called for one hour of preparation and breaks for every six hours processing orders, implying desired time per order of 1.16 person-hours). Interviews suggested that service personnel worked unreported overtime that accounted for most of the discrepancy:

... 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.

Direct observation corroborated these statements. The estimation shows that at the start of our simulation, the employees were devoting the required time to each customer. The most important result of the partial model estimation, however, 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 a time constant (τ_d) of about 19 weeks. But there is no evidence of any upward revisions in desired time per order when work pressure is low ($\tau_u \approx \infty$), 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 mind-set endure even in times of low work pressure.

Of the 34 model parameters, functions and initial conditions, we estimated 14 econometrically and another 4 directly from the data. We obtained 12 parameters either through interviews or direct observation. Four parameters, all related to work intensity and its effects, were not active during the period when data were available. For these we used the best estimate available from the existing literature. Table 2 lists all parameters and non-linear relationships, their values, and sources.⁶

4. Results

4.1. Historical fit of the model

The derivation of the model structure and parameters from observed micro-decisions and physical flows, and the ability of partial model structure to replicate data series with plausible parameters constitute a test of the model's structural validity. Furthermore, the policies estimated for the decision-makers to show that they behave rationally, albeit not optimally, within the existing incentive system (Morecroft 1985). Another test, however, is the ability of the full model to replicate historical behavior. 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, table 3).

The Mean Absolute Percent Error (MAPE) between the simulated and actual variables is less than 2% for 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 the historical series of orders processed is in part due to the fact that throughput is strongly determined by customer orders, and overtime, time per order, and hiring allowed capacity to keep pace with demand. 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 capable of tracking the overall behavior of the system variables but is not very reliable 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 the period covered by the simulation 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 six 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 down, training requirements fall. As the recently hired employees gain experience they become more effective 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 an overshoot in capacity, resulting in low work intensity.

5. Analysis

5.1. Simulations from initial equilibrium

The disequilibrium in the historical case provides a good test of the model but to understand the range of scenarios under which there might be erosion in service quality we simulated the LC model from initial equilibrium with independent random variations in customer orders and absenteeism. The driving series for the simulations were generated as pink noise (Richardson and Pugh 1981) with the same mean, variance and autocorrelation spectrum as the historical series. Figure 4 shows the first hundred weeks of a typical simulation.

Simulations showed that employees absorb small increments in demand and absenteeism⁷ by reducing time per order (B1 in figure 5) and increasing work intensity (B2). The reduction of time per order, while enabling an immediate increase in throughput, also erodes the internal service standard –the desired time per order (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 the labor requirements (B3). The reduction of 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 (R2). Despite initial equilibrium *and* stationary demand, the simulations consistently showed erosion of the service standard. In five hundred simulations the erosion rate of desired time per order over 200 weeks was, on average, 3.1% per year ($p=0.00$).⁸

Much of the observed erosion of the service standard can be explained by the lack of upward adjustment of desired time per order discussed in the estimation section ($\tau_u \approx \infty$). We found, however, that even minor asymmetries in the time constants of the standard formation process are capable of generating significant erosion rates. In simulations with a ten percent difference between the upward and downward time constants for the adjustment of desired time per order ($\tau_u = 1.1 * \tau_d = 20.7$ weeks) the service standard eroded at an average rate of 0.5% per year ($p=0.03$).⁹ With the estimated parameter and stationary demand the erosion of quality is not significantly different from zero when adjustments are fully symmetric ($\tau_u = \tau_d = 18.8$ weeks): average erosion rate = 0.3% per year ($p=0.15$). However, plausible changes in parameters cause significant quality erosion even with symmetric adjustment. 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 ($p=0.00$). Cutting normal tenure to 200 weeks, a value consistent with the low unemployment expected after the recession in the UK at the time of the study, causes average quality erosion 0.5% per year ($p=0.04$), even without demand growth.

5.2. Response to work pressure.

To illustrate how the interactions among the three responses to work pressure – increase service capacity (sc), increase work intensity (wi), and reduce time per order (TPO) – generate the erosion of the service standard, the LC model was initialized in equilibrium and tested, without noise, with a 10% step increase in customer orders. Figure 6 shows the change of throughput achieved by each of the responses, as well as the change of throughput resulting from the permanent erosion of the service standard. The combination of responses is effective in immediately increasing

throughput by 10%. However, the timing and strength of these responses are noteworthy. First, the initial reduction of TPO is almost twice as aggressive as the increase in WI. We found that workers under pressure to increase output are much more willing to cut the time they devote to each customer and only reluctantly work longer hours. At the operating point, the elasticity of work intensity to schedule pressure is 0.37, while the elasticity of time per order is 0.64. 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 in the LC during the time of the study. Consequently, the formation of quality standards is exclusively driven by employee’s perception of quality ($\omega_s=1$), and the effect of quality pressure on time per order was not statistically significant.¹⁰ Of the 15 loan officers interviewed, all but one mentioned they reduce their effort to sell additional products and document transactions in times of high work pressure. The weak response of quality pressure and the resulting willingness to cut time per order is consistent with the emphasis the monitoring system (and employees) place on processing customer orders the same day they arrive.

Second, whereas employees’ responses – TPO and WI – are instantaneous, the adjustment of SC takes 25 weeks to peak. Management has the instruments to perceive changes in labor productivity accurately and quickly ($\tau_{pe}=6.7$ weeks). Performance metrics are available on a weekly basis, but are summarized and analyzed at the end of the month. To smooth out the high frequency variations in customer orders, management adjusts the estimation of required service capacity with an average lag of 4 months ($\tau_r=18.8$ 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 4), and is consistent with management’s imperative to control costs. Once labor is authorized it takes on average seven months for the hiring process to bring a new employee into the LC ($\tau_h=29.9$ weeks). Furthermore, we found rookies to be only a third as productive as experienced personnel and requiring one year to achieve full productivity. The combination of cautious hiring policies, hiring delays, and long training requirements mean service capacity is slow to react to changes in

demand. Thus temporary variations in work pressure must be accommodated through 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 model increased its throughput 10% by reducing the internal standard of customer service by 5.4% and increasing service capacity 4.1%. The long time it takes the system to reach equilibrium, due mainly to the long training period, explains how the accumulation of random variations in work pressure generates a prolonged drift towards low quality.

The elasticity and lag of the mechanisms discussed above are summarized in table 4. The right side of table 4 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 provide 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 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 operational metrics for service quality prevents them from realizing the costs of eroding the service standard.

5.3. Implications

Does the erosion of service quality matter? Since customer service expectations adjust to past performance, it could be argued that a reduction in service standards could indeed represent a productivity gain and an effective cost reduction strategy. The downward adjustment of service quality, however, necessarily implies a transitional dissatisfaction; customers will adjust to a lower service expectation only after having experienced what they would consider a bad service

interaction. 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 time per customer on sales of business loans (£/week) achieved by the LC. Despite the large variance in the sales data, time per order is a significant predictor of loan volume. The 4.1% reduction of the service standard experienced during the period for which data were available translates into 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.

5.4. Policy Analysis

In this section we use the model to explore policies to maintain service quality without compromising the organization's ability to respond to demand fluctuations. All simulations for policy analysis were performed from equilibrium, under the expected scenario for NHLHC: 3% per year growth in demand and average employee tenure of 4 years ($\tau_a^*=200$ weeks), and with asymmetric time constants for the adjustment of desired time per order ($\tau_{ii}=1.5*\tau_{id}$). Except where noted, all other parameters had the values reported in table 2. Variations in demand and absenteeism were introduced via pink noise with same parameters as the historical series. In the base case simulation (no intervention), the service delivery system maintained the average delivery delay close to the stated goal, but the quality standard eroded at an average rate of 2.38% per year. Results are reported in table 6.

Expediting the adjustment of capacity. Since the erosion of the internal service standard occurs when work pressure is high, one obvious recommendation is to ensure that service capacity is acquired before the erosion of the service standard takes place. Service capacity can be obtained faster by having a more responsive hiring process. To test this policy we reduced the time to adjust labor and the hiring delay by 50%. The policy had a limited impact, reducing the quality rate to 2.12% per annum, 11% less than the base case.

Another strategy to increase the responsiveness of service capacity is to hire employees with higher initial effectiveness or to accelerate their learning process. A faster learning curve is critical in high-growth industries, where large number of rookies can overwhelm a service organization.

Unfortunately, these options are rarely available in high contact services that require job specific knowledge. In the expected scenario for the NHLC, because of the relatively slow growth rate, this strategy had a negligible impact on the erosion rate even when coupled with aggressive hiring.

Reducing the effect of work pressure. The positive feedback driving the erosion of the service standard operates through work pressure. Isolating service personnel from any perception of backlog or required throughput – including managerial pressure – might break the erosion loop. Eliminating variations in work pressure would ensure a consistent allocation of time to each customer but it would also reduce the center's flexibility to absorb the random variations in customer orders through overtime. To maintain a reasonable delivery delay, management would then have to carry excess capacity or have access to a flexible reserve capacity to absorb those variations.

A more realistic policy is to distribute the response to work pressure more evenly between overtime and corner cutting. This could be done by reducing the flexibility of the service standard – through standardization and documentation of the service delivery process – or by increasing the relative attractiveness of overtime; either by creating high empathy with customers or increasing overtime compensation. Standard service delivery processes are, by definition, not appropriate for high-contact services requiring customization, hence our policy analysis focused on the relative strength of the responses to work intensity. To test this policy we reduced the effect of work pressure on time per order by 50%, almost equating it to the response of work intensity. Although more successful than the rapid acquisition of service capacity, the policy had limited success sustaining the internal quality standard. When combined with the rapid acquisition policy, the policy to reduce the effect of work pressure on time per order cut the service standard erosion rate by 37% (∇ work pressure in table 6). Reducing the effect of work pressure on time per order also had the expected

consequence of increasing the delivery delay. While the average delivery delay throughout all simulations increased only by 1%, in certain periods it peaked at 7% over the stated goal.

Although the main lever to stop the erosion of the service standard is to eliminate the effects of work pressure, the data from the LC show that removing work pressure is not enough to upgrade the desired service level. Excess capacity prevents further erosion, but it does not reverse the decline of the standard. To generate upward pressure on the standard it is also necessary to maintain high quality pressure and operationalize its effects on the delivery process.

Creating quality pressure. In the absence of high professional standards – such as health care providers – management needs to take an active role in the formation of the service standard.

Creating quality pressure requires management to become aware of the implications of a lower service standard – lost sales, rework, and customer defections – inform employees about those opportunity costs, and articulate clear expectations for service quality. Even though NH loan officers reported some discomfort with their performance (evidence of quality pressure), it was not possible to identify any effects of quality pressure on the formation of the service standard, even during periods of low work pressure. To test a policy of strong quality pressure we allowed the internal quality standard to be influenced by management's quality goal ($\omega_s=0.5$), and assumed the resulting quality pressure had a strong effect on the time per order ($f_{pt}=e^{1.0p}$). We found that the effect of quality pressure on TPO has to be strong, relative to the effect of work pressure ($f_{wt}=e^{-0.64w}$), because customers rarely complain when service is close to their expectation. For quality pressure to be effective it has to amplify the small deviations reported through customer feedback.

Although the creation of quality pressure significantly reduced the erosion rate of the desired TPO (a 72% reduction from the base case), it was not enough to stop it all together. Only when all tested policies – faster acquisition of service capacity, reduced effect of work pressure on time per order, and creation of quality pressure – were implemented simultaneously did the erosion subside. The combined policy resulted in a 2% increase in the average delivery delay, with peaks up to 13%

higher than the stated goal. If critical, delivery delay performance could easily be improved by increasing work intensity, currently peaking at 108%, through greater use of overtime.

6. Concluding remarks

We have shown that service quality erodes, even under stationary demand, due to a reinforcing loop that arises from intendedly rational decisions by each actor in a service setting. Employees, in an effort to meet their throughput goals, absorb small variations in workload by reducing the time spent with each customer and task, and increasing work intensity. The reduction in time per customer, while enabling an immediate increase in throughput, also erodes the internal service standard. 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 requirements. The reduction of service capacity further increases the workload so service personnel are forced to cut the time per customer still more. We have also shown that the reduction of service standards had implications beyond the obvious costs of poor quality and that it has a direct and significant impact on the revenue generation capability of our research site.

The erosion of service standards in high-contact services is the result of the relative intensity of the available responses to work pressure – reducing time per order, increasing work intensity, and increasing service capacity – and the absence of a fixed objective standard. The relative intensity of the responses are 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, and a significant but slow learning curve reduce the speed at which service capacity can be acquired. The specifics no doubt vary from industry to industry, e.g., 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, we expect the structure that can lead to quality erosion to be common throughout the service sector.

The tendency towards erosion of quality standards is not limited to high-contact services. For example, on-line trading and other Internet businesses are currently facing unexpectedly high rates of demand growth. Although e-commerce transactions are standardized, the popular press is filled with stories about organizations that, under work pressure, fail to provide support, i.e., customize the service interaction, when something goes wrong or as customer needs change. 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.

¹ The original formulation of the model (Oliva 1996) also considered capital stocks and their technological content through a CES production function. However, for most ranges of reasonable parameters (including the parameters for 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.

² The effective labor fraction (e) is constrained to be non-negative to control for cases where rookies require more supervision than their initial effectiveness ($\eta > e$) and rookies outnumber the senior personnel ($L_r > L_e$).

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). The data from our research site do not enable us to discriminate between the two formulations. Zangwill and Kantor (1998) examine the relationships among different formulations for learning; see also Argote (1990).

⁴ The original formulation (Oliva 1996) also considered constraints on the growth rate for the labor force and layoff policies, but they had no impact on the dynamics described in this paper and hence are omitted here.

⁵ Because the average workweek for the period where data was available was always within 10% of the standard workweek (35 hrs.), the fatigue feedback was not active. Employee experience mix and its effects on productivity were estimated independently with data from 6/93 to 5/95.

⁶ The documented model is available at <http://www.people.hbs.edu/roliva/research/service/esq.html>.

⁷ The normalized standard deviations (σ/μ) of customer orders and the non-absent fraction of service capacity series were less than 4%.

⁸ P values report significance levels, under one-tailed tests, for H_0 : erosion rate = 0.

⁹ All simulations were done with the parameters estimated for the LC with the exception of the reported change. All simulations were initialized in equilibrium, and the historical variations from the driving variables introduced, via pink noise, at week 10. We report the average annualized erosion rate after 210 simulated weeks in a sample of 500 simulations; p values report significance levels, under one-tailed tests, for H_0 : erosion rate = 0.

¹⁰ 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. However, when probed about the actions implemented in response to the indicators, the manager of the LC admitted that the information was neither reliable nor useful. Furthermore, none of the lending officers interviewed were aware of the instrument nor how branch managers evaluated their performance.

References

- , “American Customer Satisfaction Index,” *American Society for Quality* (1998a), <http://acsi.asq.org/>.
- , “U.S. Products get better markets than service,” *Quality*, 37, 1 (1998b), 18.
- Argote, L. and D. Eppler, “Learning Curves in Manufacturing,” *Science*, 247(1990), 920-924.
- Baumol, W.J., “Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis,” *American Economic Rev.*, 57, June (1967), 415-426.
- Baumol, W.J., S.A. Batey Blackman and Wolf, E.N., *Productivity and American Leadership*, MIT Press, Cambridge, MA, 1991.
- Boulding, W., A. Karla, R. Staelin and V.A. Zeithaml, “A Dynamic Process Model of Service Quality,” *J. of Marketing Research*, Feb. (1993), 7-27.
- Broh, R.A., *Managing Quality for Higher Profits*, McGraw-Hill, New York, 1982.
- Chase, R.B., “The Customer Contact Approach to Services: Theoretical Bases and Practical Extensions,” *Operations Research*, 29, 4 (1981), 698-706.
- Cyert, R. and J. March, *A Behavioral Theory of the Firm*, Prentice Hall, Englewood, NJ, 1963.
- Darr, E.D., L. Argote and D. Eppler, “The Acquisition, Transfer and Depreciation of Knowledge in Service Organizations: Productivity in Franchises,” *Mgmt. Sci.*, 41, 11 (1995), 1750-1762.
- Einhorn, H.J. and R.M. Hogarth, “Behavioral Decision Theory: Process of Judgment and Choice,” *Annual Rev. of Psychology*, 32(1981), 53-88.
- Farber, B.A. (Ed.) *Stress and Burnout in the Human Service Professions*, Pergamon, New York, 1983.

- Gummesson, E., *Quality Management in Service Organizations*, International Service Quality Association, New York, 1993.
- Hogarth, R.M., *Judgment and Choice: The Psychology of Decision*, Wiley, New York, 1980.
- Homer, J.B., "Worker Burnout: A Dynamic Model with Implications for Prevention and Control," *System Dynamics Rev.*, 1, 1 (1985), 42-62.
- Jarmain, W.E. (Ed.) *Problems in Industrial Dynamics*, MIT Press, Cambridge, MA, 1963.
- Kahneman, D. and A. Tversky, "The Psychology of Preferences," *Scientific American*, 246(1982), 160-173.
- Koepp, S., "Why is service so bad? Pul-eeze! Will somebody help me?," *Time*, Feb. 2, 1987, 46.
- Lant, T.K., "Aspiration Level Adaptation: An Empirical Exploration," *Management Sci.*, 38, 5 (1992), 623-644.
- Levinthal, D. and J.G. March, "A Model of Adaptive Organizational Search," *J. of Economic Behavior and Organization*, 2, 4 (1981), 307-333.
- Lewin, K., T. Dembo, L. Festinger and P.S. Sears, "Level of Aspiration," in J.M. Hunt (Ed.), *Personality and the Behavior Disorders*, Ronald Press Co., New York, 1944, 333-378.
- McKinsey Global Institute, *Service Sector Productivity*, McKinsey&Co., Washington, DC, 1992.
- Mills, P.K., *Managing Service Industries*, Ballinger, Cambridge, MA, 1986.
- Mobley, W.H., *Employee Turnover: Causes, Consequences and Control*, Addison-Wesley, Reading, MA, 1982.
- Morecroft, J.D.W., "Rationality in the Analysis of Behavioral Simulation Models," *Management Sci.*, 31, 7 (1985), 900-916.
- Oliva, R., *A Dynamic Theory of Service Delivery: Implications for Managing Service Quality*, PhD Thesis, Sloan School of Management, MIT, Cambridge, MA, 1996.
- Richardson, G.P. and A.L. Pugh, *Introduction to System Dynamics Modeling with Dynamo*, MIT Press, Cambridge, MA, 1981.
- Schneider, B., "Service Quality and Profits: Can you have your cake and eat it, too?," *Human Res. Planning*, 14, 2 (1991), 151-157.
- Schneider, B. and D.E. Bowen, "Employee and Customer Perceptions of Service in Banks: Replication and Extension," *J. of Applied Psychology*, 70(1985), 423-433.
- Schneider, B., J.J. Parkington and V.M. Buxton, "Employee and Customer Perceptions of Service in Banks," *Admin. Sci. Quarterly*, 25, 2 (1980), 252-267.
- Senge, P.M., "Catalyzing Systems Thinking within Organizations," in F. Masaryk (Ed.), *Advances in Organizational Development*, Ablex, Norwood, NJ, 1990, 197-246.
- Senge, P.M. and J.D. Sterman, "Systems Thinking and Organizational Learning: Acting Locally and Thinking Globally in the Organization of the Future," *European J. of Oper. Res.*, 59, 1 (1992), 137-150.

- Shiba, S., A.K. Graham and D. Walden, *A New American TQM: Four Practical Revolutions in Management*, Productivity Press, Cambridge, MA, 1993.
- Simon, H.A., *Models of Man: social and rational; mathematical essays on rational human behavior in a social setting*, Wiley, New York, 1957.
- Sterman, J.D., "Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models," *Dynamica*, 10, Winter (1984), 51-66.
- Strandvik, T., *Tolerance Zones in Perceived Service Quality*, PhD Thesis, Swedish School of Economics and Business Administration, 1994.
- Theil, H., *Applied Economic Forecasting*, North-Holland, New York, 1966.
- Thomas, H.R., "Effects of Scheduled Overtime on Labor Productivity: A Literature Review and Analysis," Source Document, Pennsylvania State University, University Park, PA, 1993.
- Tornow, W.W. and J.W. Wiley, "Service Quality and Management Practices: A Look at Employee Attitudes, Customer Satisfaction and Bottom-Line Consequences," *Human Res. Planning*, 14, 2 (1991), 105-116.
- van Horn, R.L., "Validation of Simulation Results," *Management Sci.*, 17, 5 (1971), 247-258.
- Weisberg, J., "Measuring Worker's Burnout and Intention to Leave," *Quality of Working Life*, 15, 1 (1994), 4-14.
- Zangwill, W. and P. Kantor, "Toward a Theory of Continuous Improvement and the Learning Curve," *Management Sci.*, 44, 7 (1998), 910-920.
- Zeithaml, V., L. Berry and A. Parasuraman, "The nature and determinants of customer expectations of service," *Academy of Marketing Sc.*, 21, 1 (1993), 1-13.
- Zeithaml, V., A. Parasuraman and L. Berry, *Delivering Quality Service: Balancing Customer Perceptions and Expectations*, Free Press, New York, 1990.

Table 1 Estimates for the adjustment of time per order (eq. 4, 5, 11, 12, 13, 14)

	Estimate	95% Conf. 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_{ij}	814,000	327,000 ∞

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

Table 3 Historical fit June 1994-May 1995

	MAPE	Theil's Bias	Inequality Unequal Variation	Statistics Unequal Covariation	R ²	N
Desired labor	0.009	0.109	0.257	0.633	0.740	52
Total labor	0.008	0.026	0.143	0.830	0.747	52
Time available	0.009	0.019	0.255	0.725	0.938	50
Orders processed	0.003	0.000	0.299	0.701	0.990	50
Time per order	0.017	0.033	0.095	0.872	0.799	50
Work intensity	0.017	0.060	0.154	0.784	0.635	50

Table 4 Responses to work pressure and consequences

	Response		Consequence		
	Elasticity	Avg. response lag ⁽¹⁾	Affected State Variable	Avg. response lag	Avg. perception lag
TPO response	-0.64	0	Desired time per order	$\tau_{tdb}=18.7$	$\infty^{(2)}$
WI response	0.37	0	Fatigue	$\tau_{fc}=3.0$	$\tau_{pc}=6.7$
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_{tdb}), 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.

Table 5 Effect of time per order on sales[†]

$$\begin{aligned} \text{SQRT}(\text{Business Loan Volume}) &= a_0 + a_1 \text{ TPO} \\ &= -719 + 778 \text{ TPO} \\ \text{S.E.} &\quad (304.7) \quad (287.1) \end{aligned}$$

n=50; 10 left-censored observations (volume<0);

$\chi^2_1=7.02$ (p<0.008)

[†] Using TOBIT estimation. Results are also significant (p<0.01) under OLS.

Table 6 Policy analysis – Desired Time per Order erosion rate

Policy	Parameter changes	DTPO erosion ⁽²⁾	p value	avg. delivery delay (wks)
Base case		-2.38%/yr	0.000	0.100
Fast capacity acquisition	$\tau_l=6, \tau_h=15$	-2.12%/yr	0.000	0.100
Low effect of work pressure	$f_{wt}=e^{-0.32w}$	-1.72%/yr	0.000	0.101
▽ Work pressure	$\tau_l=6, \tau_h=15, f_{wt}=e^{-0.32w}$	-1.49%/yr	0.000	0.101
Δ Quality pressure ⁽¹⁾	$f_{pl}=e^{1.0p}, \omega_e=0.5$	-0.66%/yr	0.000	0.102
Work & Quality pressure ⁽¹⁾	$\tau_l=6, \tau_h=15, f_{wt}=e^{-0.32w}, f_{pl}=e^{1.0p}, \omega_e=0.5$	0.29%/yr	0.992	0.102

(1) The parameters driving managers' perception and formation of expectations of service quality, although not active in the NHLC, were judgmentally set based on interviews: $\omega_m=0.5, \tau_{qm}=12, \tau_{em}=52, \tau_{qc}=52$.

(2) The reported rates are the average annualized erosion rate after 210 simulated weeks over 500 simulations; p values report significance levels, under one-tailed tests, for the H_0 : erosion rate=0.

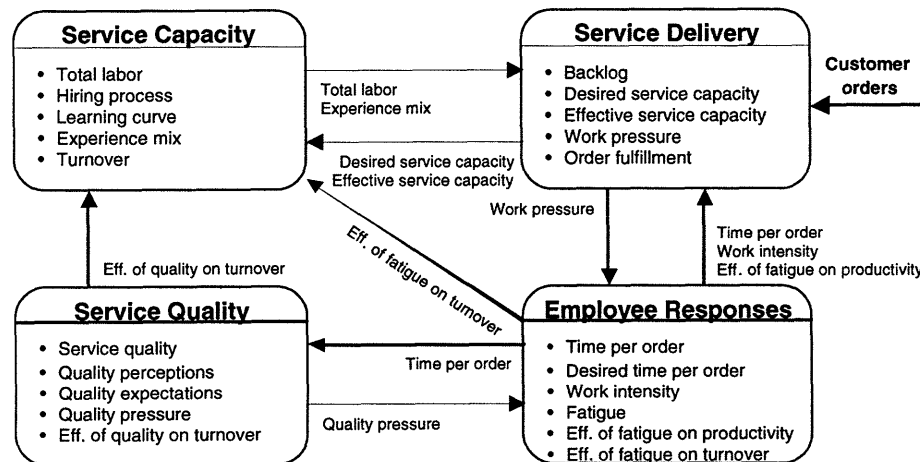
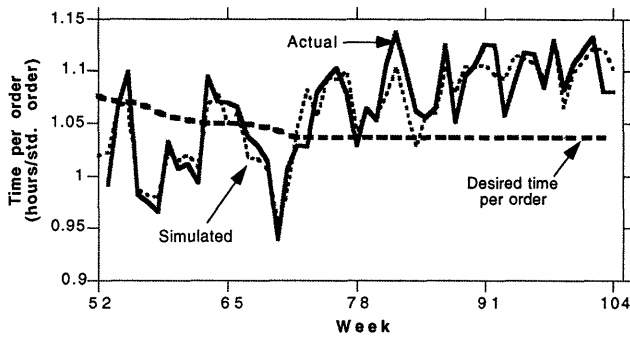
Figure 1 Model structure overview

Figure 2 Time per order (partial model estimation)



**Summary Statistics for Historical Fit
Time per Order**

n = 50

R^2	0.828
Mean Abs. Pct. Error	0.015
Root Mean Square Error	0.019
Bias	0.000
Variation	0.047
Covariation	0.953

Figure 3 Comparison of simulated and actual data

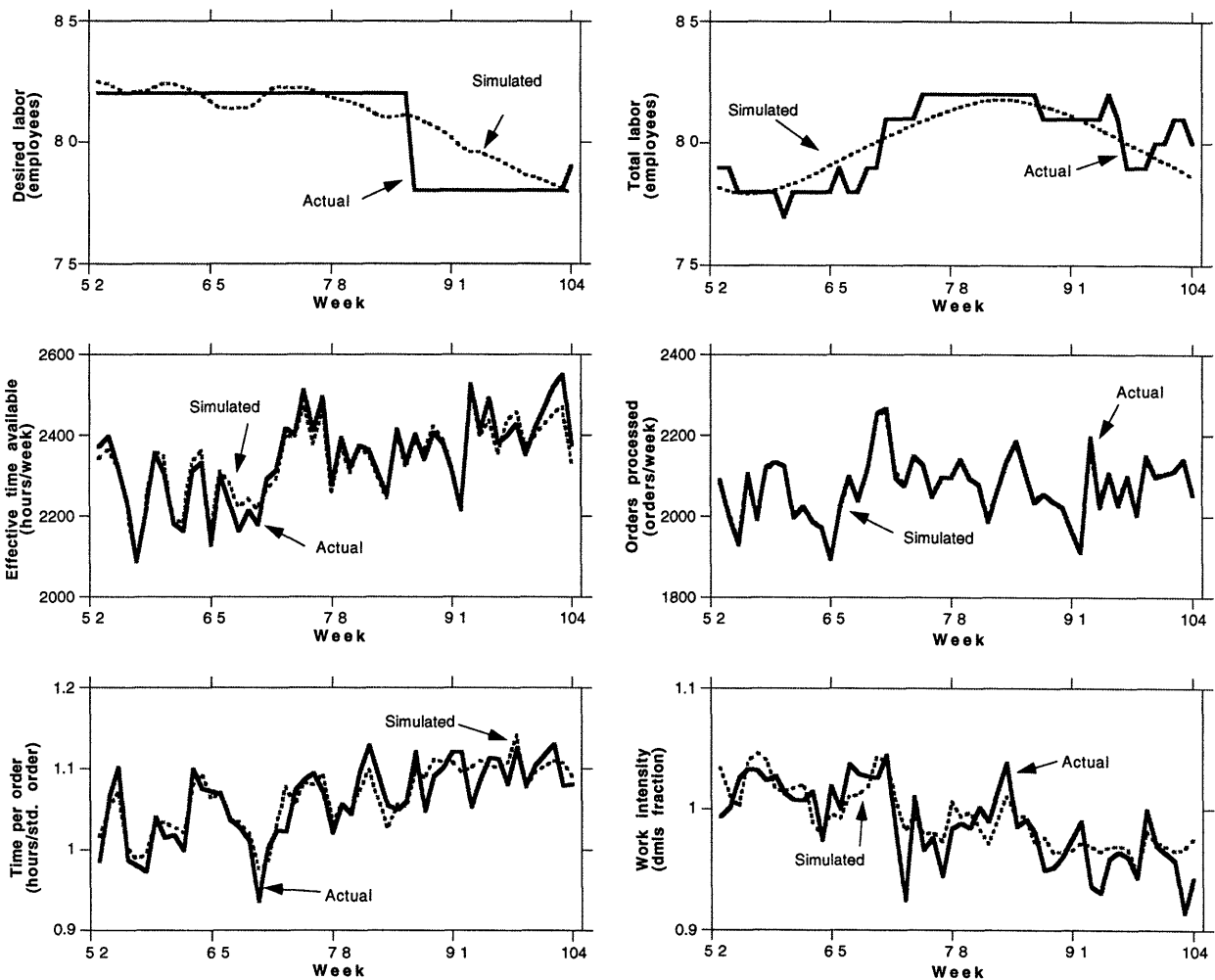
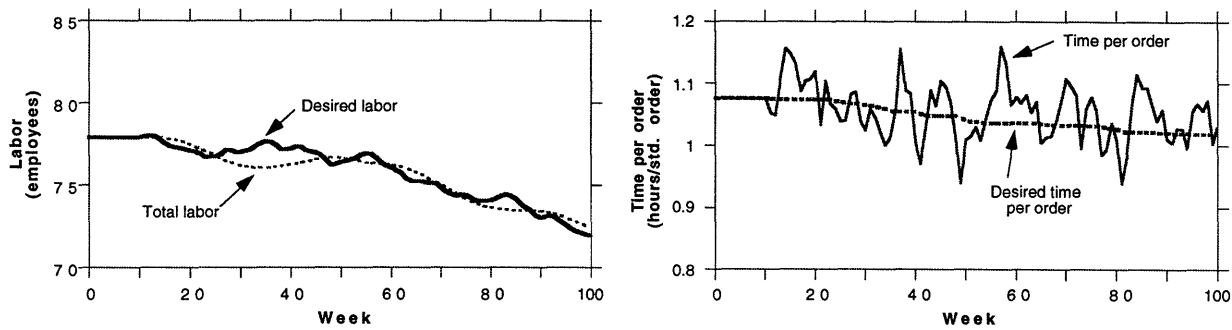


Figure 4 Response to random variations in customer orders and absenteeism[†]



[†] Both series were stationary and generated as pink noise with the same mean, standard deviation and time constant for the autocorrelation spectrum as the historical series: customer orders = pink noise (2071, 77.5, 1); absenteeism = pink noise (0.165, .032, 2).

Figure 5 Feedback structure of erosion of service standard

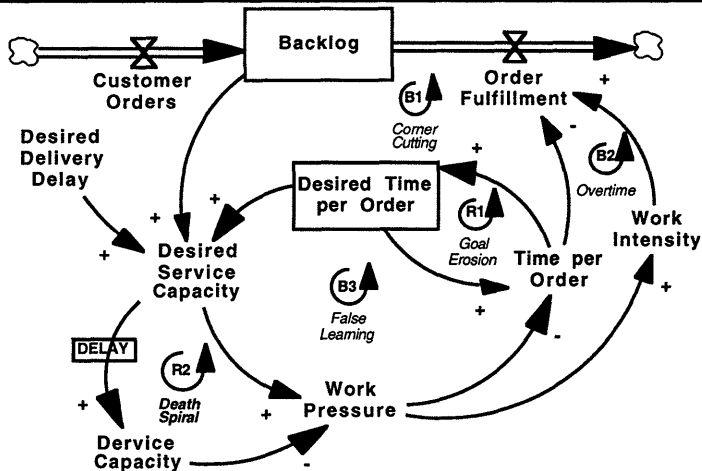


Figure 6 Response to a 10% increase in demand

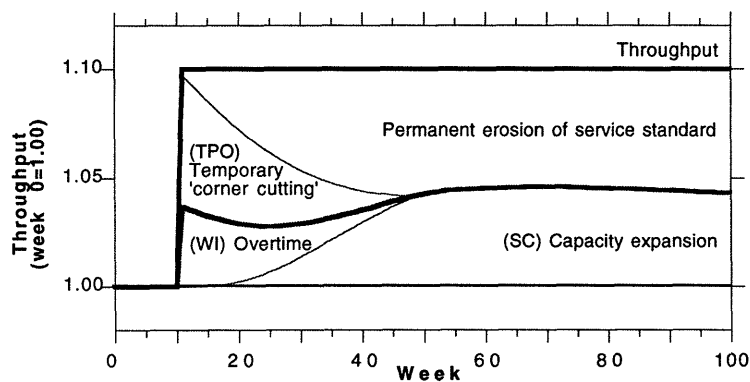


Table 2 Parameters and sources for service model

Parameter		Value		Source
Service delivery				
τ_f	Min. time required to process an order	0.1	week	set based on observations
λ	Desired delivery delay	0.1	week	set based on interviews and stated goals
f_{wt}	Effect of workload on time per order	$e^{-0.64w}$	dmls	estimated to fit past data on time per order
τ_{ti}	Time for upward adj. of time per order	813,564	week	estimated to fit past data on time per order
τ_{td}	Time for downward adj. of time per order	18.8	week	estimated to fit past data on time per order
f_{wi}	Effect of workload on work intensity	$e^{0.37w}$	dmls	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 $F_e \in [1.14, 2]$	$1-0.5F_e$	dmls	set based on previous studies
f_{fa}	Effect of fatigue on attrition $F_a \in [1, 2]$	$1-0.2F_a$	dmls	set based on previous studies
Service capacity				
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_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_v	Time to cancel vacancies	1.0	week	judgmentally set based on interviews
τ_e	Time for experience	12.0	week	judgmentally set based on interviews
τ_a^*	Time for attrition	401.0	week	estimated to fit past data on attrition
ϵ	Relative effectiveness of rookies	0.35	dmls	judgmentally set based on interviews
η	Fracc. of exp. personnel for training	0.05	dmls	judgmentally set based on interviews
Perception of service quality				
τ_{qe}	Time for employees' perception of quality	4.0	week	judgmentally set based on interviews
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
f_{qa}	Effect of quality on attrition	1.00	dmls	judgmentally set based on interviews
f_{pt}	Effect of quality pressure on time per order	$e^{0.00p}$	dmls	estimated to fit past data on time per order
Formation of quality expectations				
ω_e	Weight for employees' quality expectation	1.0	dmls	judgmentally set based on interviews
τ_{ee}	Time for employees' quality expectation	26.0	week	judgmentally set based on interviews
ω_m	Weight for mgmt's quality expectation			not active in base simulation
τ_{em}	Time for mgmt's quality expectation			not active in base simulation
ω_c	Weight for customers' service expectation	1.0	dmls	set as <i>a fortiori</i> assumption
τ_{ec}	Time for customers' service expectation			not active in base simulation
μ	Customer's service expectation reference	1.16	hours/order	estimated to fit past data on time per order
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
F_e	Fatigue for effect on effectiveness	1.00	dmls	set based on historical data
F_a	Fatigue for effect on attrition	1.00	dmls	set based on historical data
T^*	Desired time per order	1.07	hours/order	estimated to fit past data on time per order
E	Perception of labor effectiveness	0.78	dmls	estimated to fit past data on desired labor
S	Employees' perception of quality	0.95	dmls	estimated to fit past data on time per order

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