A DYNAMIC THEORY OF SERVICE DELIVERY:
IMPLICATIONS FOR MANAGING SERVICE QUALITY

by

Rogelio Oliva Pue

BE, Industrial and Systems Engineering
Instituto Tecnológico y de Estudios Superiores de Monterrey, México (1985)

MA, Systems in Management
University of Lancaster, UK (1988)

Submitted to the Sloan School of Management
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1996

Signature of Author ....
Sloan School of Management
April 8, 1996

Certified by ....
John D. Sterman
Professor of Management Science
Thesis Supervisor

Accepted by ....
Birger Wernerfelt
Chairman, Ph.D. Committee
Sloan School of Management

A Dynamic Theory of Service Delivery:
Implications for Managing Service Quality

by

Rogelio Oliva Pue

Submitted to the Sloan School of Management
on April 8, 1996 in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

ABSTRACT

This thesis presents a theory of service delivery grounded in the operations management, marketing and human resources literature that articulates an endogenous explanation for the erosion of service quality. The theory states that service quality cannot be measured and tested in as straightforward a manner as quality can be measured and tested in manufacturing. The difficulty in developing quality metrics has biased service businesses to focus on controlling measurable variables -- typically, expenses and work flows -- while under-investing in the more intangible factors of service capacity and service quality. In the long-term, this strategy can result in mediocre levels of service quality, poor customer satisfaction, low customer loyalty, and high turnover of service personnel. Ultimately, this result can change the cost structure of service delivery by shifting the major cost component from operating expenses to costs of poor quality. The new cost structure causes poor financial performance that leads management to tighten the control of expenses and work flows, creating a vicious cycle of eroding service quality.

A system dynamics model articulating the service delivery theory was empirically validated via calibration to match the structure and behavior of a service center. Overall, the model -- calibrated with information about micro-decisions and internal policies in the service center -- provided a reasonable explanation of the operational flows and the macro-behavior of the main indicators of the research site, thus increasing confidence in the structural and replicative validity of the model.

The results from the calibration process conformed to the hypothesized relationships and behavioral components of the proposed theory of service delivery. Simulations showed that, as predicted by the theory, the structural elements of the research site -- policies and physical flows -- predispose its service quality to erode even when customer orders and labor are stable.

The findings from the validation exercise were used to generate insights and derive policy recommendations for managing service quality in high-contact service settings. Finally, the model structure was used to link structural parameters of service settings to the problematic dynamics observed in the service industry.

Thesis Supervisor: John D. Sterman
Title: Professor of Management Science

To Susie

ACKNOWLEDGMENTS

Among the many people who have contributed to the work that culminates in this dissertation, the following deserve special acknowledgment.

First I thank my wife Susie for her constant support and encouragement throughout the entire process, and for being there to share the joys and frustrations of this work. While I have been doing this work she has created a home in which our two daughters can grow and learn surrounded by love. It is to Susie that I dedicate this dissertation with love and gratitude.

I am indebted to my parents who taught me to challenge myself and, despite the physical distance, have shared the ups and downs of writing a thesis.

I was fortunate to have a thesis committee that shared a deep interest in my work and held me to a high standard while providing guidance and encouragement. John Sterman has been instrumental to the successful completion of the thesis. He guided me throughout the process with academic rigor and integrity, knowing exactly when to push and when to encourage. Through insightful comments on the many drafts and countless meetings, John made many improvements on the structure of the model and the overall thesis structure. Gabriel Bitran's course on services first brought my attention to the difficulties of managing service delivery. His emphasis on validation issues and managerial implications contributed significantly to the quality of the thesis. Peter Senge introduced me to the Hanover Insurance model and taught me how to convey system insights to managers during my two years as research assistant at the Organizational Learning Center. He continued to guide my research offering suggestions on testing strategies, model validity issues, and new ways to present this work.

I thank the people at NatWest Bank for their commitment to this work and for allocating the resources to make it possible. I greatly appreciate the extensive efforts of Peter Crispin, Neil Jones, David Rogers and Kerry Scott, as well as the openness and trust extended by the employees of Nelson House.

A great source of support and encouragement were my fellow SD doctoral students — they were there and listened when I needed to talk. Anjali Sastry set an admirable example as she guided me through the doctoral program. Ed Anderson, Tom Fiddaman, Nittin Joglekar, Nelson Repenning and Scott Rockart helped with formulation issues and strategies on how to present this work. Special thanks are due to Liz Krahmer for the innumerable drafts she edited and the outstanding support she provided me with as friend and colleague.

In Mexico I would like to thank Leonel Guerra, Antonio Guzman, Heriberto Leyva and Cuauhtémoc Olmedo for helping me discover the joys of academic life and setting me on the path that led to MIT.

Finally, my thanks go to Boston's Tree of Life City Church. It was from this community, through their loving example and support, that I learned the important lessons that I take away from my stay in Cambridge.

Table of Contents

List of Figures

List of Tables

1. Introduction

§1.1. Motivation

The economy of industrialized countries, and that of the US in particular, is shifting towards services (Fuchs, 1968; Gershuny, 1978; Quinn and Gagnon, 1986; Summers, 1985). Not only do developments in the agriculture and manufacturing sectors cascade down to generate employment in the service sector as they create a need to market and distribute the goods (Cohen and Zysman, 1987), but 'value-added services' have been identified as a mechanism to ensure customer satisfaction or gain competitive advantage for companies in the manufacturing sector (Band, 1991; Lash, 1989; Vendermerwe, 1993). Furthermore, internal services (e.g., accounting, information systems, personnel, maintenance, R&D, etc.) are increasingly significant to manufacturing industries as their processes become more complex and automated. The fraction of the United States National Income that is generated in the service sector has grown from 58% in 1959 to over 71% in 1988, and by 1985 almost three out of every four employed Americans worked in the service sector (US Department of Commerce, 1992)[1].

The expanding importance of services, together with the finding from the Profit Impact of Marketing Strategy program (PIMS) that perceived quality is the main determinant of market share and profitability (Buzzell and Gale, 1987), has generated an enormous interest to better understand and manage the delivery of high-quality services.

---

[1] Included in the Service sectors are the following categories of the National Income and Product Accounts: transportation, utilities, communications, wholesale trade, retail trade, finance, insurance and real estate, government, and other services.

Unfortunately, most of the research work done in the quality of goods arena has proven inadequate for understanding service quality. Fundamental differences in the way services are produced, consumed and evaluated make the lessons from the literature on quality and consumer behavior inoperative in a service context (Zeithaml, Parasuraman and Berry, 1990).

Researchers from the operations management, human resources and marketing have dedicated considerable efforts to explore the main determinants of service quality. Most research has focused on the following areas:

- applying management science models and total quality management (TQM) principles developed for the manufacturing environment into service settings (Drewes, 1991; Hostage, 1975; Ingle and Ingle, 1983; Rosander, 1989; Sasser, Olsen and Wyckoff, 1978; Wyckoff, 1992),

- classifying services according to various typologies to facilitate their management (Chase, 1981; Haywood-Farmer, 1988; Lovelock, 1983; Schmenner, 1986),

- exploring the impact of human factors in service performance and profitability (Schlesinger and Zornitsky, 1991; Schneider, 1991; Schneider, Parkington and Buxton, 1980; Tornow, 1991; Tornow and Wiley, 1991; Ulrich, Halbrook et al., 1991), and

- explaining customers' perceptions and expectations of service quality (Boulding, Staelin et al., 1992; Gronroos, 1984; Maister, 1984; Zeithaml, Parasuraman and Berry, 1990).

Although some integrated frameworks of service delivery and service quality have been articulated (Bitran and Lojo, 1993a; Bitran and Lojo, 1993b; Heskett, Jones et al., 1994; Schlesinger and Heskett, 1991), most of the evidence available for the relationships proposed in these frameworks is fragmented. No multidisciplinary studies testing the implications of the integrated findings from operations management, marketing and human resources are available in the literature. Because of the lack of an integrated and accepted framework of service delivery, guidelines for managers to understand and control the inherent biases and conflicts in the delivery of service quality are also absent from the literature. The work presented in this dissertation was motivated by both of these shortcomings. The purpose of this research is to develop and test an integrated theory of

service delivery capable of generating insights into the challenges of managing service quality.

§1.2. Background

The starting point for this research was a dynamic hypothesis of the interactions between service capacity and service quality that was articulated in the context of a multiple-year system dynamics study with Hanover Insurance Company (Kim, 1989; Senge, 1990a; Senge, 1990b; Senge and Lannon, 1990; Senge and Sterman, 1992). Concisely, the dynamic hypothesis states that service quality cannot be measured and tested in as straightforward a manner as quality can be measured and tested in manufacturing. The difficulty in developing quality metrics has biased service businesses to focus on controlling *measurable* variables – typically, expenses and work flows – while under-investing in the more *intangible* factors of service capacity and service quality. In the long-term, this strategy can result in mediocre levels of service quality, poor customer satisfaction, low customer loyalty, and high turnover of service personnel. Ultimately, this result can change the cost structure of service delivery by shifting the major cost component from operational costs to costs of poor quality. The new cost structure causes poor financial performance that leads management to tighten the control of expenses and work flows, creating a vicious cycle of eroding service quality.

In the six years since the original theory of service delivery was developed in the insurance context, the model has been recast as a generic theory for high-contact services (Oliva, 1993b; Senge and Oliva, 1993) and the basic model has been turned into a flight simulator (MicroWorlds, 1994; Oliva, 1992; Oliva, 1993a) and used in workshops for hundreds of managers from diverse service industries. From this experience, it was speculated that the findings from the Hanover Insurance case are applicable to a wider set of service settings. Specifically, two attributes of high-contact services – the difficulty of assessing service quality and the tight coupling between service personnel and service delivery – create the context for a theory capable of explaining the reference mode of decreasing profitability while lowering operating expenses. The next section describes the approach used in this dissertation to develop and validate the integrated theory of service delivery.

§1.3. Approach and Overview

The research work described in this dissertation can be grouped into three distinct stages: 1) formalization of the dynamic theory of service delivery and its substantiation in the existing service literature, 2) empirical validation of the theory, and 3) derivation of managerial implications of the theory. The next subsections briefly describe the approach followed for each of these stages and its location within the text of the dissertation.

Formalization and Substantiation of Theory

The proposed theory of service delivery integrates findings from different disciplines that have examined the service delivery process. The theory, while being grounded in the human resources, behavioral decision theory, marketing, and operations management literature, is articulated using a system dynamics model along with a detailed account and evidence from the literature for the proposed constructs, causal linkages, and formulations that compose the theory.

A computer simulation model can be an effective tool for validating theory. First, the model formalizes the hypothesized relationships between variables creating a refutable causal model with multiple 'points of testing' (Bell and Bell, 1980; Bell and Senge, 1980). Second, it enables testing of the completeness and coherence of the proposed relationships (Sastry, 1995; Sterman, 1985b). System dynamics is appropriately suited for this integration because of its emphasis on causal linkages, the rigorous formulation of these linkages that it requires, and the ability to test, via simulation, whether the relationships defined at the micro-level are capable of generating the macro-behavior of the whole service setting.

A brief literature survey on service and service quality is presented in Chapter Two. The theory is presented with full documentation and substantiation as a system dynamics model in Chapter Three. Chapter Three also explores briefly the dynamic implications of the theory.

Empirical Validation of the Theory

Although the proposed theory describes the relationships between variables throughout the service setting, much of the evidence available for those relationships is fragmented and specific to the relationships. No full validation of all the simultaneous interactions is

yet available. In testing a complex dynamic theory, there are three validity concerns that should be addressed:

- Does the micro-structure of the model correspond to what is known about the real system?

- Do the estimated or observed relationships support the theory?

- Can the macro-behavior of the service setting be explained from the structural components of the theory?

These concerns guided a validation strategy based on calibrating the existing model of service delivery to fit the structure and behavior of a service setting. The selected service setting was a back-office center in a major British bank responsible for making loan decisions for the mass market (personal loans and credit cards) and small business accounts (sales less than £100,000 per year).

To address the structural validity issue – the extent to which the model captures the structural elements of the real system – the calibration was done through partial model estimation with immediate data sources. Data were collected through direct observation of the service delivery process, interviews with personnel responsible for the service delivery and support functions and time series of the center's operational metrics.

The other validity concerns were addressed through a suite of tests performed at the full system level. Replicative validity was tested through the historical fit of the model. The dynamic significance of the structural components was tested through sensitivity analysis. Finally, extended simulations were used to test the overall dynamic hypothesis articulated by the theory. The rationale behind calibration as a validation methodology and the results and findings from the partial model estimation and the full system tests are described in Chapter Four.

Managerial Implications

Although the model – calibrated with information about micro-decisions and internal policies in the service center – provided a reasonable explanation of the operational flows and the macro-behavior of the main indicators of the research site, not all the hypothesized relationships were corroborated by the validation process. Both the expected and surprising results were source of insights about the theory and high-contact service contexts. Specifically, the findings from the validation process were used to generate insight into the relative strength of the different responses to work pressure and

to propose a more parsimonious and empirically appealing formulation for the formation of service aspirations.

A second set of managerial implications, perhaps the most important one, was the identification of leverage points and policy recommendations for managing quality in a high-contact service setting. A validated system dynamics simulation model, with a closed causal boundary (Forrester, 1969), i.e., most of the variables are internal to the service center and under managerial control, provides an endogenous theory of the dynamics of service quality and profitability. The endogenous perspective, by making explicit the conflicts among variables, allows the model to be used to explore alternative intervention strategies for improving service quality.

Finally, to facilitate the generalization and transferability of insights, the model was taken outside the high-contact service context and its usefulness in other service settings explored. By explicitly examining the application domain of the theory – the set of structures and behaviors the theory is capable of explaining – it was possible to define a generic framework to link structural characteristics of service settings to the problematic dynamics observed in the service industry.

The three levels of managerial implications – insights about service delivery, policy recommendations for high-contact services, and generalization and transferability of insights – are presented in Chapter Five.

§1.4. Contributions

By having an empirical component – the direct comparison of a theory with the specific real world setting that is attempting to describe – this dissertation makes contributions in both the substantive and methodological domains.

§1.4.1. Substantive Contributions

The substantive contributions to the field of service quality of this dissertation go in hand with the managerial implications explained above:

First, the model resulting from this research is a coherent and internally consistent theory of the dynamics of service delivery and quality that provides an endogenous explanation of erosion of service quality. As such, the model constitutes a contribution to the understanding of the dynamics of service delivery and quality.

Second, by being the enactment of a theory, i.e., a simplification of the real world, the model permits to isolate the most significant determinants of system behavior. The ability to isolate dominant factors allows to use the model to explore alternative strategies and derive policy recommendations for managing service quality. An explicit set of policy recommendations for the management of service quality in high-contact services was developed in this dissertation.

Finally, the structure of the model was used to generate a framework to classify service delivery processes according to their dominant structural characteristics and their impact on the behavior of the service setting. A classification that links structural characteristics to behavioral expectation facilitates the transferability and generalizability of insights about the service delivery processes.

§1.4.2. Methodological Contributions

Methodological concerns address the question of how to investigate a particular subject matter. This dissertation is grounded in two academic traditions – operations management and system dynamics – with slightly different methodological principles and values for what constitutes a good theory, model, and validation strategy. In attempting to satisfy the two perspectives, the research approach followed in this work has expanded the traditional research methodologies in each of these fields.

The methodological contribution to *operations management* is two-fold. First, this work makes a contribution to the processes of modeling situations with soft behavioral variables. The use of the system dynamics modeling methodology not only allowed for the inclusion of behavioral variables in the model – decisions by managers, employees and customers – but also revealed, through simulation, the importance of these variables in the operating dynamics of the service setting.

The second methodological contribution to operations management is an approach to formally integrate interdisciplinary findings. Studies of the service provision process are inherently eclectic – they require insights from operations management, human resources, economics and marketing (Lovelock, 1992). However, the integration of the different findings and theories about the service delivery process has been difficult to achieve. The dissertation proposes a way to integrate these multidisciplinary findings and derive their implications. The system dynamics model developed achieves a formal integration that allows for explicit testing of implications and generation of policy recommendations.

The research is also makes two methodological contributions to the *system dynamics* field. First, the dissertation articulates and documents a model validation strategy that explicitly addressees the validation concerns from the management sciences perspective. Although the validation strategy was developed with the idea of testing a preexisting model in a real world situation, the same strategy could be used to test dynamic hypotheses in a traditional system dynamics intervention.

The second methodological contribution for system dynamics is a methodological strategy to explore the transferability of insights from the model into different service settings. One of the long standing claims of system dynamics has been that of generalizability, i.e., the creation of a common frame of reference to capture the characteristics of a system and make them transferable to other settings (Forrester, 1961). The kernels of transferable knowledge in the system dynamics field have been captured as generic structures – "relatively simple models of dynamic processes that recur in diverse settings and that embody important management principles" (Senge, 1985, pg. 791). Forrester's claim "... that about 20 such general, transferable, ... cases would cover perhaps 90 percent of the situations that managers ordinarily encounter" (1993, pg. 210) testifies to their perceived importance in the development of the field.

The system dynamics literature on validation has focused on construct and internal validity (Judd, Smith and Kidder, 1991), but has not fully explored the dimension of external validity – "the extent to which one can generalize the results of a research to the populations and settings of interest" (*ibid*. pg. 28). Without addressing the issues of external validity, it is impossible to make the generalizability claim, and, therefore, difficult for generic structures to become part of mainstream management theory. The approach followed in this dissertation is presented as the first steps for a methodological strategy to address the external validity issue of system dynamics models.

2. Service Delivery and Service Quality

§2.1. Introduction

The previous chapter gave a clear sense of the high profile that the service sector has in the US economy and how important manufacturers consider services to be for the delivery of product packages and customer satisfaction. The service sector of the economy has not always been held in such high regard. Classical economic theorists regarded services as non-productive activities. The only source of wealth was the accumulation of tangible assets (capital), and, since the output of service was ephemeral, services were considered non-productive work and outside the main concerns of economics and management (Delaunay and Gadrey, 1992). However, during the first half of the twentieth century, it became obvious that more and more people were being employed by activities that were unrelated to the extraction of transformation of physical goods. By measuring the growth of real national product, income and consumption in three broad sectors – primary (agricultural), secondary (manufacturing), and tertiary (services) – Clark (1940) captured this shift of employment towards the tertiary sector and described it as the transition of an industrial economy to a post-industrial economy. The nature of this shift and its sociological implications were later explored by Bell (1973). Bell identified four attributes of the post-industrial society that shifted the perceived importance of services in the economy: (1) the post-industrial society as a tertiary (service) society; (2) the primacy of knowledge, science and technology; (3) the preeminence of the professional and technical class; and (4) the change of value systems and forms of control.

Once services were recognized as a major source of for economic growth there were strong incentives to increase the efficiency and effectiveness of service delivery processes. Management science (MS) models developed for manufacturing organizations – e.g., queuing theory, staff scheduling, site location – were applied to services. Although useful, the MS models proved to be inadequate for services.

> As useful as these methodologies are, however, they are based upon a manufacturing model of business that cannot capture the subtle interaction among the customer, delivery systems, and economics that characterizes the complex world of service management (Chase and Heskett, 1995, pg. 1717).

One possible explanation for the insufficiency of the MS models in service is the assumption brought from manufacturing that the units being processed and the stations providing the service are passive, i.e., products, batches, machines, etc. Service delivery processes, however, are different because both the servers and the units being processed are human – with psychological attributes, perceptions and expectations. Services are produced, consumed and evaluated differently from goods (Zeithaml, Parasuraman and Berry, 1990).

The purpose of this chapter is to place the research in this dissertation within the context of existing work in the service literature and the current challenges being faced by high-contact services. The next section presents a summary of the main characteristics that differentiate services from manufactured goods. Section 2.3 explores the implications of these characteristics in terms of challenges for managing high-contact services and the models that have been developed to tackle them. Finally, the chapter concludes by listing how different challenges and aspects of the service delivery process are incorporated into the theory articulated in this work.

§2.2. Characteristics of High-Contact Services

Chase (1981) introduced the 'customer-contact approach' to analyze and design service settings (Chase and Aquilano, 1989; Chase and Tansik, 1983). From this perspective, service businesses are evaluated according to *potential facility efficiency* – a decreasing function of the "degree to which the customer is in direct contact with the service facility relative to the total service creation time" (1981, p. 700). The continuum created by this indicator allows for a classification of *pure*, *mixed* and *quasi-manufacturing* service settings. Using that classification as a framework, Chase derived a set of characteristics for high-contact services (table 2.1) and explored different possible strategies for intervention by decoupling the interaction between the customer and the service center.

1. The service product is multidimensional (time, place, atmosphere) and hence its quality is in the eye of the beholder.
2. The direct worker is part of the service product.
3. Demand for the service is often instantaneous and hence cannot be stored.
4. Because production is generally customer initiated, an optimal balance between service system demand and resources is difficult to achieve.
5. Changes in the capacity of the system affect the nature of the service product.
6. The production schedule has a direct, personal effect on the consumer.
7. Only part of the service can be kept in inventory.
8. Verbal skills and knowledge of policy are usually required of the service worker.
9. Wage payments must usually be related to labor hours spent rather than output.
10. It is assumed that service system capacity is at its long run level when the service first opens.
11. A service system malfunction will have an immediate, direct effect on the consumer.
12. The location of the service system modifies its value to the customer.

Table 2.1   Characteristics of high-contact services – Some propositions
Source: Chase, 1981.

Chase implicitly assumes that there is a direct relationship between the amount of time that the server interacts with the customer and the degree of responsiveness (customization) of the service setting. Schmenner (1986) refined this classification by identifying two dimensions in the continuum from high to low customer contact. He argued that it is possible to have high-contact service settings with very limited and structured interactions (e.g., a hotel); while others would require a much more responsive and customized approach (e.g., a hospital). To capture the interaction-time dimension of the 'customer-contact approach,' Schmenner defined an indicator of labor-intensity as "the ratio of the labor cost incurred to the value of the plant and equipment" (1986, p. 21). This indicator, he argued, reflects the relative importance of the interaction with the customer. He places the two dimensions in a two-by-two matrix (table 2.2) from which different challenges and strategies are identified for managers of each category of service setting.

| | | Degree of Interaction and Customization | |
		Low	High
	Low	**Service Factory:** - Airlines - Trucking - Hotels - Resorts and Recreation	**Service Shop:** - Hospitals - Auto Repair - Other Repair Services
Degree of Labor Intensity	High	**Mass Service:** - Retailing - Wholesaling - Schools - Retail Aspects of Commercial Banking	**Professional Service:** - Doctors - Lawyers - Accountants - Architects

Table 2.2   The service process matrix
Source: Schmenner, 1986.

Although professional services are the archetypal case for labor-intensive and high-customization services, the label seems to exclude other types of services in which personnel skills are the main determinant of service quality, e.g., insurance claims adjusters, tax auditors, bank loan officers, salespeople of complex products, nurses, etc. The label *high-contact* services will be used to capture the essence of these service settings with high customer contact and high degree of customization.

The characteristics of the service delivery process identified by Schmenner reflect two attributes that distinguish services from manufacturing products: the difficulty of measuring the output of the service delivery process and the predominant role of service personnel in the delivery process. Services are *intangible* and *labor intensive*.

The intangibility of services creates difficulties in describing the product to customers. It is equally difficult for customers to express precisely what they desire from the product. Service intangibility and high labor intensity constitute major obstacles towards the standardization of the delivery process and meeting customers' agreed requirements. Because of the inherent variations in customers' expectations and service personnel, heterogeneity arises in the perception of delivered service quality.

Since the two dimensions represent a continuum, one possible strategy – suggested by Levitt (1972; 1976) and later re-taken by Schmenner (1986) – is to redesign the service delivery process to fit the traditional manufacturing model, i.e., shift the service delivery process to the upper-left corner of table 2.2. This is the approach taken by the fast-food industry. However, high-contact services, by definition, require labor intensity and customization to satisfy customer expectations.

A third attribute that differentiates services from manufacturing products is that service production and consumption are simultaneous. The instantaneous nature of the provision of service and its consumption results in an interaction between server and consumer enabling either party to influence the quality of the service. *Inseparability* of service delivery and service quality means that there is no time buffer in which checks can be made to assess whether the quality of service meets a required standard. Customers see all mistakes. Provision of the service is the 'moment of truth' and Total Quality Management's dictum "right the first time" takes on added importance.

§2.3. Challenges Introduced by Service Characteristics

Each of the these attributes – intangibility, labor intensity and inseparability – bring challenges to the managing of the service setting that limit the usefulness of operations management models developed for manufacturing. The difficulties introduced by these characteristics to the management of service delivery processes are discussed in the next subsections.

§2.3.1. Difficult Assessment and Management of Service Quality

The fact that the output of a service delivery interaction is not tangible creates the possibility for customers to rely on other indicators to determine the quality of the interaction. Furthermore, since there is no established agreement on the service to be delivered, customers walk into the service center with different levels of expectations. These expectations, in turn, shape the perceptions that customers have of the service delivery. The vast literature in service quality and the variety of models that has emerged from it can be taken as indicators of the theoretical difficulties of defining service quality and the challenges to operationalize and manage it (see Hoech (1988) and Gummesson (1993) for a review and synopsis of the literature).

The first challenge that emerges from the lack of an objective measure of service quality is that customers rely on other indicators to determine service quality. Quality is not evaluated by the customers solely in terms of the outcome of the service; they also consider the process of service delivery. This renders service quality a multidimensional construct that encompasses all aspects of service delivery. In the most comprehensive study in this area, Zeithaml, Parasuraman and Berry (1990) identified, through exploratory customer studies, five orthogonal dimensions that customers use to make their assessment of service quality: tangibles, reliability, responsiveness, assurance and empathy.

The second difficulty introduced by the lack of tangible quality metrics is a shift in the definitions of service quality towards a subjective evaluation based on customer expectations. Since there is no objective normative standard the only criteria available to evaluate service quality are measures defined by customers. According to this view the customer's assessment of service quality results from a comparison of customer's expectations to their perception of the actual service delivered (Gronroos, 1984; Maister, 1984).

The realization that service quality is based on customer expectation makes it clear that it is not possible for management to have direct access to the factors defining quality. The most articulated model of this perspective (Parasuraman, Zeithaml and Berry, 1985; Zeithaml, Parasuraman and Berry, 1990) argues that the difference between customer expectations and actual service provided cannot be managed directly but only through other 'gaps,' or discrepancies, between expectations and performance that occur in organizations. Figure 2.1 is the graphical representation of these gaps.

Gap 1   the difference between what consumers expect and what management perceives them to expect,

Gap 2   the difference between management's perceptions of consumer expectations and actual service quality specifications,

Gap 3   the difference between service quality specifications and the service delivered,

Gap 4   the difference between service delivery and what is communicated about the service to consumers, and

Gap 5   the difference between the customers' perception and their expectations of the service.
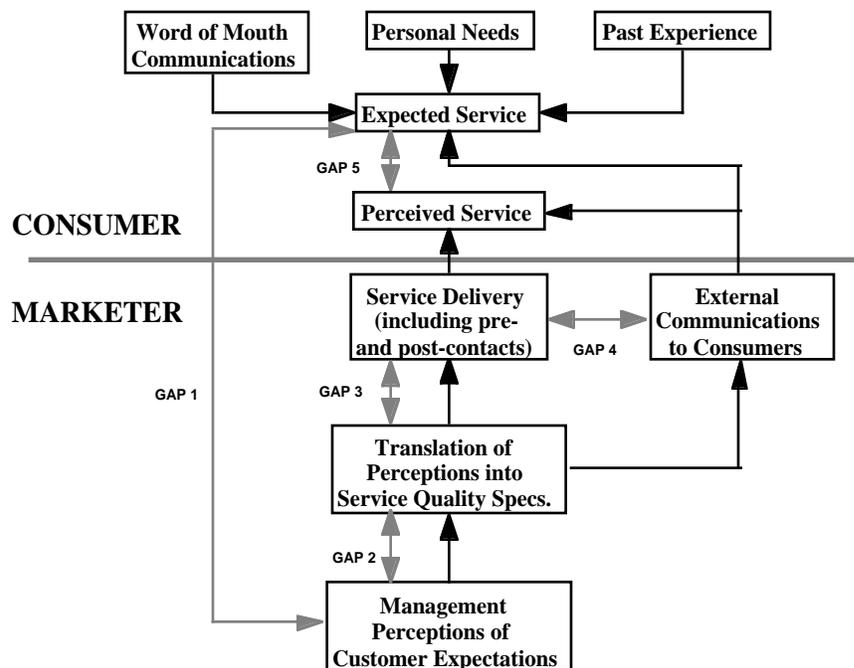
Figure 2.1   Conceptual model of service quality
Source: Parasuraman, Zeithaml and Berry (1985).

The authors of this model have identified a set of theoretical constructs and variables to assess each of the gaps and developed a set of communication and control principles to

reduce the size of, and noise generated within, each of the manageable gaps (Zeithaml, Berry and Parasuraman, 1987).

A further implication of the subjective assessment of service quality that has not been explored by the authors of the gap model is the fact that customers' expectations evolve through time, thus making service quality a moving target. For example, software users have grown more experienced over time so that what were once advanced features are now basic requirements, and new capabilities are required. Evolving customers' expectations point to the need for a dynamic process to adjust service specifications to match the ever-changing customers' needs (Aranda, Fiddaman and Oliva, 1993).

§2.3.2. Pressures from Limited Productivity Growth

Because of the tight coupling in the service delivery process between output and personnel (propositions 2, 3, 4 and 9 in table 2.1), significant gains in productivity cannot be expected through capital substitution in high-contact services. As an extreme example of a labor-intensive service, Baumol (1967) suggests imagining the effects of trying to reduce the personnel required to play a half hour horn quintet to anything less than two and a half person-hours.

Obviously, no single service is purely generated in the 'customer contact' phase; thus, some gains in productivity could be expected in the 'backroom' operations (Bitran and Lojo, 1993a). Table 2.3 presents a comparison of the annualized rates of productivity change from 1958 to 1976 in selected services as grouped by the Standard Industrial Classification codes. The industry groups are ranked according to their labor intensity – calculated as the ratio of deductions for depreciation to employees' compensation in a year. Although some authors have argued that the service sector has had significant productivity gains (Quinn and Gagnon, 1986; Shelp, 1988), It is clear from the regression coefficients in table 2.3 that the service industries with high capital-labor ratios are the ones with higher productivity gains.

	Capital-Labor Ratio†	Productivity Growth*
Pipeline transportation	0.9982	7.9%
Electric Utilities, Gas, Sanitation	0.6293	3.8%
Motion Pictures	0.4587	2.7%
Automobile repair	0.3878	2.8%
Communications, except radio and TV	0.3463	5.6%
Air Transportation	0.3277	4.9%
Transportation Services	0.2695	-1.7%
Railroad transportation	0.2194	5.2%
Radio and TV Broadcasting	0.1665	-0.4%
Hotels and lodging places	0.1410	1.8%
Local transit and intercity buses	0.1315	0.2%
Amusement and recreation services	0.1077	0.8%
Truck transportation	0.1002	1.6%
Banking	0.0982	0.0%
Credit agencies and financial brokers	0.0597	-2.8%
Personal and repair services	0.0532	1.0%
Retail Trade	0.0519	1.4%
Insurance	0.0452	1.5%
Wholesale trade	0.0383	3.1%
Educational Services	0.0045	1.2%

Prod growth	n=20	F-Ratio=14.84
$R^2 = 0.42$	**Coefficient**	**t-stat**
Constant	0.3917	0.6422
C-L Ratio	7.0695	3.8530

Table 2.3   Productivity changes by service type

† Estimates based on 1967 data on depreciation deduction and compensations of employees .
Source: *Statistics of Income, Corporation Income Tax Returns.* Internal Revenue Service, 1967,
          Table 1 and *National Income and Product Accounts of the United States,* US Department
          of Commerce, 1992, Table 6.2B-6.2C.
* Annual average rates of change based on least squares method (1958–76).
Source: *Time Series Data for Input-Output Industries: Output, Price and Employment.* US
          Department of Labor, 1979. Bulletin 2018.

The lower productivity growth in high-contact services generates some operational difficulties for the service delivery process. These difficulties were first articulated by Baumol (1967) through a simplified model of the economy in which every activity was placed into either a 'stagnant' or 'progressive' sector. The sectors in this model are characterized by different rates of productivity growth as determined by the role that labor plays in the activity. "In some cases labor is primarily an instrument—an incidental requisite for the attainment of the final product, while in other fields of endeavor, for practical purposes the labor itself is the end product" (Baumol, 1967, pg. 416). With a

very limited set of assumptions[1], he demonstrated that the unbalanced growth of productivity will cause the cost of a unit of output in the stagnant sector to grow persistently and cumulatively, while it would remain constant in the progressive sector[2]. Increasing unit cost translates into financial pressure on the firms of the stagnant sector – the so-called 'cost disease' (Harker, 1995).

Baumol further concluded that if the ratio of the output of the two sectors was to be held constant (a balanced demand for the output of both sectors), more people would be required to work in the stagnant sector, assuming a constant quality of the output. He did realize the implications of his findings for service quality:

> "... there is one implicit underlying danger that should not escape the reader: the inherent threat to quality. Amateur activity has its virtues ... But in a variety of fields it offers a highly imperfect substitute for the highly polished product that can be supplied by the professional. Unbalanced productivity growth, then, threatens to destroy many of the activities that do so much enrich our existence, and give to others over into the hands of the amateurs. These are dangers which many of us may feel should not be ignored or take lightly." (1967, pg. 422).

Erosion of service quality is but another form of economizing and increasing productivity (Stanback Jr., 1979). The data in table 2.3 clearly shows that not all the service industries belong to Baumol's stagnant sector. However, the high labor-intensity industries clearly show a slower growth of productivity, exposing them to the financial pressures or the potential erosion of service quality predicted by Baumol.

§2.3.3. Performance Based on Attitudes and Perceptions[3]

The third challenge introduced by service characteristics is product of the proximity between service personnel and customers during the service transaction. Service personnel are brought into the service transaction because they are an intrinsic element of the delivery process (high labor intensity). Consumers, on the other hand, are required to participate because of the instantaneous consumption of service and the required co-production of services. This structure brings employees and customers physically, organizationally and psychologically close – thus blurring the boundary between employees and consumers (Schneider and Bowen, 1985).

---

[1] a) Costs other than labor can be ignored, b) wages in the two sectors go up and down together and c) wages will rise as rapidly as output per person-hour in the sector where productivity is increasing.

[2] Baumol's model has since been validated with empirical observations (Baumol, Blackman and Wolff, 1985).

[3] This section draws from the synthesis of this literature done by Tornow (1991) and Schneider (1991).

Schneider, Parkington and Buxton (1980) showed that when employees report a service imperative in their bank branch, customers report that they receive high service quality. The service imperative was defined by Schneider et al. by a set of questionnaire scales, completed by employees, regarding the following issues: management emphasis on service through rewards or service goals, adequacy of service personnel in terms on numbers and capacities, adequacy of supplies and equipment, and emphasis on retention of customers in daily activities. These results were later extended (Schneider and Bowen, 1985) to show that when employees describe positive Human Resource (HR) practices in their work place, i.e., employees have a "positive work attitude," customers report higher quality services. Furthermore, the study showed that customer's intentions to switch to another bank were predictable based on employees' perceptions of service quality delivered to customers, and that employee turnover intentions were predictable based on customers' perception of service quality.

The positive relationship between employees' and customers' perceptions and attitudes, and their link to turnover intentions has been replicated and well documented in the HR literature (Tornow and Wiley, 1991; Wiley, 1991). Further studies have also shown a positive relationship between employees' perceptions of organizational commitment to quality and their attitudes towards service delivery, and moderate evidence linking employees' perceptions and attitudes to unit profitability (Schlesinger and Zornitsky, 1991; Ulrich, Halbrook et al., 1991).

§2.4. Research Scope

Although the driving mechanisms for the challenges described above are well understood and documented in the service literature, little work is has been done to understand the effects of these driving forces acting simultaneously in a service setting. Since the purpose of this research is to develop and test an integrated theory of service delivery capable of generating insights into the challenges of managing service quality, the difficulties expressed in the previous sections are explicitly accounted for in the proposed theory of service delivery.

Two additional concerns help define the scope of this research. First, I seek to capture the *dynamic* interactions among the structural elements and agents – managers, service personnel and customers – in the service delivery process. Second, I have an interest to generate *endogenous* explanations for the behaviors and conflicts observed in the service industry. A dynamic perspective ensures that the theory will capture the evolution of

customer expectations and the learning that takes place as organizations continuously adjust to meet those expectations. The focus on endogenous explanations, i.e., most variables are internal to the service center and under managerial control, is to facilitate the transferability of insights to different service settings and allow the model to be used by managers to explore alternative strategies.

Specifically, the dynamic theory of service delivery articulated in the following chapter incorporates the following elements of the service delivery process and service quality:

- Regarding the assessment and management of service quality, the theory explicitly accounts for the dynamic behavior of customers' expectations, management's quality standards and the actual service quality delivered. In terms of the gap model, the proposed theory articulates the perceptual biases, delays, and feedback mechanisms governing gaps 1, 2, 3 and 5.

- Regarding the pressures from the unbalanced productivity growth, the theory explicitly models technological innovation (productivity gains), the financial pressures to control costs and the responses that employees and management might have to those pressures.

- Finally, regarding the human factors affecting the service delivery process, the theory accounts for employees' perception of service quality and customer satisfaction, the skill level and experience that customer-facing personnel have, the effects of fatigue on productivity and turnover and the effects of employee satisfaction on turnover.

A detailed articulation of the dynamic theory of service delivery with evidence from the literature for the proposed constructs, causal linkages, and formulations is given in the next chapter.

3. A Dynamic Theory of Service Delivery

§3.1. Introduction

In this chapter I propose a causal theory to capture the major characteristics of the service delivery process and articulate the endogenous argument for erosion of service quality presented in Chapter One. The theory is presented as a system dynamics model with evidence from the literature for each of the hypothesized causal links.

A brief description of the historical development of the emerging theory of service delivery is first presented, followed by the definition of new constructs and assumptions necessary to articulate a theory for the high-contact services. Section 3.4 provides an overview of the theory, focusing on the underlying structures in a service setting and its dynamic implications. Section 3.5 formally articulates the theory as a system dynamics model, defining the scope and boundary of the model, and giving a detailed account and evidence for the proposed constructs, causal linkages, and formulations that compose the theory. The chapter concludes with simulation results of the base model linking observed behavior to specific model structure.

§3.2 The Hanover Insurance Case

The theory of service delivery guiding the present research has been developed over several years. The first articulation of the theory emerged in the context of a multiple-year study with Hanover Insurance Company (Kim, 1989; Senge, 1990a; Senge, 1990b; Senge and Lannon, 1990; Senge and Sterman, 1992). That study focused attention on the rising costs of claims settlements and litigation (figure 3.1), and the declining overall financial health of the property and casualty insurance industry.
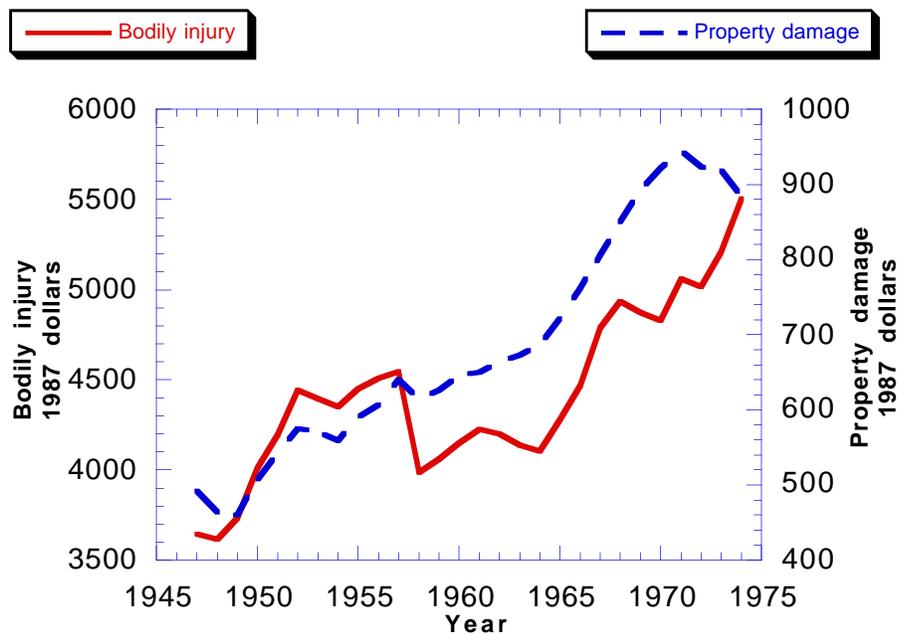
Figure 3.1   Average claim cost – auto-liability insurance

Source: Insurance Information Institute as reported by Standard & Poor's Industry Surveys (1963 – 1991).
Values in constant 1987 dollars using the Implicit Price Deflator for Final Sales to Domestic
Purchasers. *National Income and Product Accounts of the United States,* US Department of
Commerce, 1992. Table 7.2.

During the last 50 years, there has been a rising trend in 'loss ratios' (settlement costs and litigation costs relative to premiums) and a decline in 'expense ratios' (operating expense relative to premiums) for the property and casualty insurance industry (figure 3.2). Within the industry, rising settlement and litigation costs are often blamed on external factors, such as the high number of lawyers in the US, increasing litigiousness of society, the tendency for juries to side with victims rather than 'big business' insurers, and increasing risks born of technological complexity (Huber, 1987). Additionally, one might interpret the falling expense ratios as evidence of increasing productivity and management innovation. The study illuminated internal sources of the problems and suggested a different explanation for the falling expense ratios.

The central hypothesis that emerged from the study is that rising settlement costs and falling expense costs are causally related: there has been a long term trend of underinvestment in service capacity that has resulted in erosion of quality of investigation, negotiation, and customer service, resulting in rising costs of settlement and litigation. Moreover, the savings in expenses have been more than offset by the increases in costs of poor quality. The consequent long-term increase in total costs and erosion of profitability have led to increasing focus on expense control and productivity

(normally defined as customers served per service person per time unit), thereby reinforcing underinvestment in service capacity.
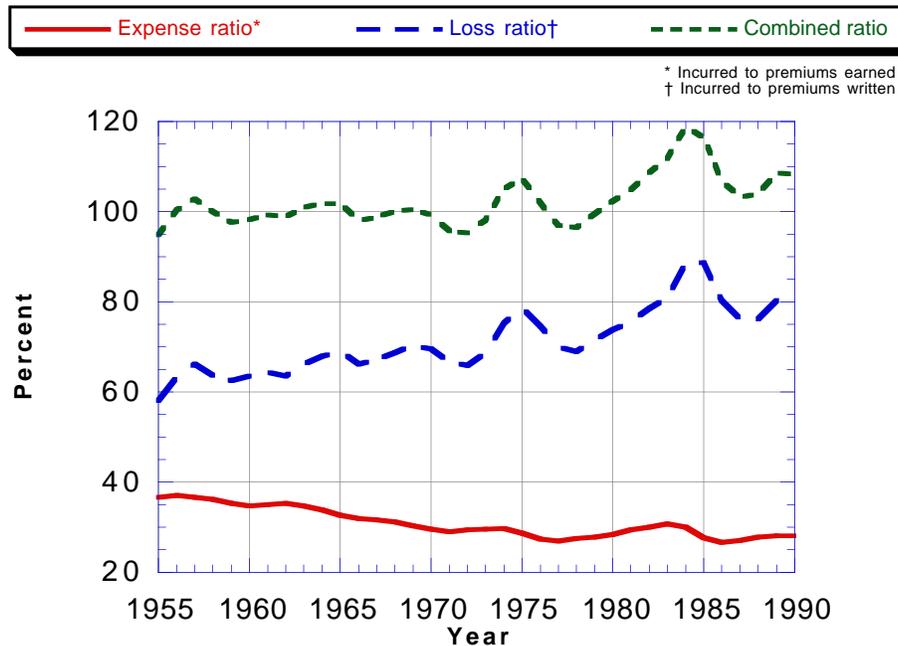
Figure 3.2   Underwriting ratios for stock property – casualty companies
Source: A.M. Best Co. as reported by Standard & Poor's Industry Surveys (1963 – 1991).

A good theory, according to the system dynamics paradigm, links observable macro, i.e., system-wide, patterns of behavior to micro-level decision making (Forrester, 1979; Morecroft, 1983). The first efforts to develop the above hypothesis focused on showing how the underinvestment dynamic could emerge from interactions among goals, norms, performance measures, and pressures that managers in the insurance industry could identify in their own experience. A team led by the Vice President of Claims in the sponsor company collaborated to develop a system dynamics model showing how established management practices and policies could produce underinvestment and rising total costs. The process whereby the initial model was developed is described by Senge:

> The key to the hypothesis lay in distinguishing two classes of performance measures: 'production standards' and 'fuzzy standards.' Production standards are measures such as 'production ratio' and 'pending ratio,' which indicate whether current claims pending are settled at a rate commensurate with the inflow of incoming claims. The production standards are relatively easy to measure, are understood by everyone in the business, and send out clear immediate warning signals when they become out of balance. The fuzzy standards include quality of investigation, file quality, effective oversight of litigation and subrogation (recovery of costs from other insurers), and service quality. The fuzzy standards are difficult to measure. Though there is widespread appreciation that the fuzzy standards are important, the team felt that there is usually considerable uncertainty as to how well a claims office is doing on the 'fuzzies.' (Senge, 1990a, pg. 216-217).

Initial simulation tests of the SQ-SC model showed that a focus on production standards could be problematic under times of high customer demand. In particular, two simulation tests showed that it was impossible to distinguish two different adjustment mechanisms that might operate in response to an increase in incoming claims if one focused only on production measures (Senge, 1990a, pg. 221-222). In one case, production measures readjust to acceptable balances because of increasing adjuster capacity. In the other case, they readjust because of eroding fuzzy standards. Thus, if management tracks only the production measures, it is impossible to know what is happening at a deeper level: desired production levels may be maintained only because of eroding quality. In a simulation where incoming claims grow steadily, there is a rising volume of claims settled per adjuster along with a steady decline in fuzzy standards, a pattern that matches qualitatively the historical trend of falling expense ratios and rising loss ratios.

Subsequent field trials with claims managers revealed consistencies in their decision making in simulation experiments that corroborated the theory. In particular, Kim (1989) studied claims managers working with a management flight simulator in which they made decisions regarding hiring, setting production targets, and setting targets for settlement size – a surrogate for service quality. He found that "although the company professedly emphasized pursuing high quality standards, the behavior in the games showed that controlling expenses dominated people's actions." (1989, pg. 333). Moissis (1989) also showed that simple decision rules used consistently throughout the simulation were capable of surpassing the performance of most managerial teams.

Although there has been limited assessment of the effects of the learning lab on claims management practices, according to the senior managers who facilitated the lab, there were impacts on local managers particularly in the area of a heightened sensitivity to interconnections between expenses, production, quality, service capacity, and total cost (Cavaleri and Sterman, 1995). Cavaleri and Sterman also report some behavioral changes as a result of the learning laboratories – new metrics that emphasize service quality, and changes in organizational structure and hiring policies that increased the effective processing capacity. However, some externalities were confounded with the operational performance of the claims adjusting unit in a way that was difficult to assess the impact of those changes on the unit's business results.

§3.3. On Service Quality and Service Capacity

In chapter 2 I identified some attributes of high-contact services that make it difficult to apply traditional operations management concepts to a service setting. To cope with the difficulties that these characteristics introduce, it is necessary to define new constructs that will permit the articulation of testable causal theories in the service domain. In this section I delineate such constructs and their assumptions based on the characteristics of service processes and service quality. Descriptions of the specific formulations and metrics for these constructs will be given in §3.5.

Because of the *inseparability* of service quality and service delivery, it is desirable to capture in a single construct both the throughput component of service delivery, i.e., fulfillment of customer orders, and the quality of the service provided. On the other hand, the *intangibility* of service quality argument states that quality is not only a function of the service provided, but also a function of multidimensional customer expectations. Although employees can normally have a sense of customers' satisfaction, service quality is intrinsically subjective and difficult to assess and monitor.

To address the issue of *inseparability* service quality has been defined as a function of the time allocated per customer order – a proxy for the degree of attention and care that servers are providing – and customers' expectations (see eq. 44 in §3.5). To deliver higher service quality more time is required from the service provider – clearly the case for the claims adjusters where, to do a better investigation and maintain more complete and accurate records, they had to spend more time on each claim. The relationship between time allocated per order and service quality holds for the high-contact service settings where customer interactions are important and perceived service quality suffers if customers feel rushed by their servers.

The assumption that time per order is the main driver of service quality is consistent with Mills' equation of service quality to server productivity (1986, pg. 127), and a commonly made claim that "the most important component of a service is personnel" (Broh, 1982, pg. 174) – see also Rosander (1989, pg. 43).

To be congruent with the assumption of quality being a function of the time allocated per order, the traditional definition of service capacity – time available for processing orders – needs to be expanded from person-hours to include personnel's skills, attitudes, and efforts, as well as the technological content of the service delivery process. The proposed

formulation incorporates these issues by calculating the nominal service capacity through a production function of labor and capital. The production function estimates the nominal service capacity based on the availability of these factors, the balance between them, and the technological content of capital (see eq. 1 in §3.5). Effective service capacity is obtained by modifying the nominal service capacity by labor effectiveness – a function of personnel's skills, work intensity, and enthusiasm (see eq. 38 in §3.5).

Consistent with the *intangibility* assumption, quality is considered to be a summary metric of customer satisfaction that is difficult to assess. No knowledge is assumed about the relationship between the service center operating parameters and customer satisfaction. In principle, quality could be measured along the five dimensions proposed by Zeithaml et al. (1990) – tangibles, reliability, responsiveness, assurance, and empathy. However, because these elements are often correlated, and to simplify the modeling effort, quality is represented as a scalar.

The challenge then is to identify a metric that is easier to obtain and monitor than the subjective service quality. Time per order is hypothesized to be such a metric. Although the proposed metric for service capacity contains some 'fuzzy' elements – worker skills, attitudes, and effort – these elements are internal to the service center and in a way easier to measure than the intrinsic customer expectations. The model explores the implications of 'poor quality' – reimbursements, cost of processing complaints and rework, lost sales, etc. – out of the more tangible metric of time per order. Table 3.1 presents a summary of the proposed functions for these constructs.

service quality	$= f$(time per order, customer expectations)
time per order	$= f$(orders to process, service capacity)
service capacity	$= f$(nominal service capacity, labor effectiveness)
nominal service capacity	$= f$(capital, labor, technology)
labor effectiveness	$= f$(skills, attitudes, effort)

Table 3.1   Service quality constructs

The proposed constructs view the service delivery process from a very particular distance. The constructs are not detailed enough to look at specific dimensions of service quality nor the expectations and satisfaction of individual customers. On the other hand, the constructs are not that removed from the operations of the service delivery process as

to ignore its internal structure or the pressures that employees face when dealing with customers[1].

With the expanded interpretations for service capacity and service quality, it is possible to present a theory of the service delivery process capable of explaining the issues presented in §3.2. The theory will be articulated first as a dynamic hypothesis, i.e., a qualitative description of the behavior of the system and its underlying causes. After this generic overview, I will describe the theory's constructs and hypothesized causal arguments as a system dynamics model.

§3.4. Dynamic Hypothesis

A dynamic hypothesis is an explanation of how structure is causing behavior. In this section, I will concentrate on the main causal structures of the model and their implications for the behavior of the system. A full description of the model equations, the rationale for the key formulations, and evidence for the main causal links will be presented in §3.5.

The service delivery model represents a service center where customers enter the system and, after a waiting-time, are served by the center's employees. Service capacity is a function of the personnel in the service center, their effectiveness, the capital available – infrastructure – and the technology embodied in the capital; and it is measured in homogeneous service hours. Each customer order requires a minimum time to be fulfilled, and the quality of the transaction is assumed to be a function of time allocated to it compared to the customer's expectations. Management of the service center defines throughput goals, quality goals, and the acquisition of service capacity. Managerial decisions are represented by variables in bold-italic characters in figures 3.3–3.7[2]. Full consideration to those policy formulations will be given in the detailed description of the model structure.

The basic order flow in the service center is captured by two rates de-coupled through a stock that represents the backlog of customers awaiting service. The service backlog is augmented by the inflow of customer orders, and reduced by the rate of orders processed (see figure 3.3).

---

[1] This 'distance' is consistent with the view taken by System Dynamics studies (see Forrester, 1961, pg. 96).
[2] These diagrams do not portray all the information flows contained in the model.
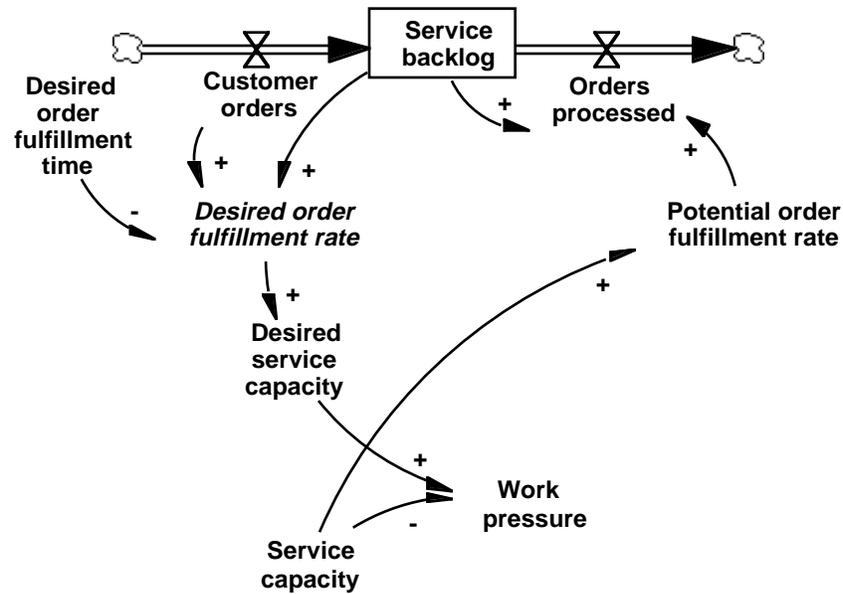
Figure 3.3   Causal loop diagram. Service throughput

The size of the service backlog is controlled through the managerial order fulfillment decision that seeks to maintain the average delivery delay (service backlog/orders processed) equal to a desired goal. Work pressure is defined as the gap between the service capacity required to deliver the desired order fulfillment rate and the actual service capacity.

The balancing loops in figure 3.4 represent mechanisms through which the system reacts to an increase of work pressure – increasing the work intensity (**B2**), reducing the time allocated per order (**B3**) or obtaining additional resources through investment (**B4K**) or hiring (**B4L**) (Senge and Sterman, 1992). While the decisions to increase work intensity and reduce the time per order are typically controlled by the service personnel; the acquisition of additional service capacity is controlled by corporate headquarters or the management of the service center. All of these responses seek to increase the potential order fulfillment rate thus creating various balancing loops that drive service backlog to its desired level.

Figure 3.4   Causal loop diagram. Responses to work pressure

The reinforcing loops **R1**, **R2**, and **R3** (see figure 3.5) capture the effects of sustained work intensity as a response to an increase in work pressure (**B2**). If work intensity is maintained above normal levels, employees will eventually become fatigued thus reducing their effectiveness (**R1**). If the strategy is sustained even longer, the effects of fatigue will translate into burnout and, eventually, increased turnover. A high turnover rate not only reduces the amount of total personnel available to perform the service (**R2**) but also reduces the accumulated experience in the service center (**R3**).

Figure 3.5   Causal loop diagram. Effects of sustained work intensity

Loops **R2** and **R3** give a detailed account of the 'cycle of failure in service' identified by Schlesinger and Heskett (1991). "High turnover reinforces the wisdom of decisions to minimize efforts in selection, training, and commitment-building activities ... this cycle produces indifferent attitudes toward customers and poor service" (*ibid*., pg. 17). In the formulation chosen here, management's emphasis on training and retaining experienced personnel is captured in experienced personnel. If turnover increases, labor effectiveness falls as the experience base is diluted. Low labor effectiveness eventually triggers higher work intensity thus reinforcing the turnover rate. As Schlesinger and Heskett note, these processes can be reversed to produce better service quality. It is this structure that makes the pursuit of higher quality an attractive goal, as it becomes a self-reinforcing virtuous cycle that increases productivity and reduces cost (Deming, 1982).

By adding management's response to financial pressures, i.e., cost control – **B5L** and **B5K** in figure 3.6 – it is possible to see how a service setting that already has a natural tendency for quality erosion falls into the trends observed in the insurance data of decreasing profitability while lowering operating expenses. Managers' attempt to keep costs under control – reducing capital investment and hiring – constrains the effective service capacity. A reduced responsiveness in acquiring service capacity makes the system susceptible to work pressure under any increase in customer demand. Work pressure, if sustained, translates into lower delivered quality, thus increasing the costs of poor quality that further intensifies the financial pressures (**R4L** and **R4K**).

Figure 3.6   Causal loop diagram. Response to financial pressure

If the hiring decision is set to maintain total personnel at a constant level and the desired order fulfillment rate is modeled to maintain average service time equal to a target – both realistic policies under financial pressure – the loops presented heretofore are enough to generate the observed erosion of service quality, i.e., a reduction of the time allocated per order, as a response to an increase in orders (see simulation results in §3.6). Several mechanisms, however, are in place to avoid a free-fall of service quality (see figure 3.7).

Delivered quality is perceived, with some delays and biases, by employees participating in the service delivery, and by management, either through direct observation or through feedback from market research instruments. If employees perceive that their quality is deteriorating, as compared to an internally held standard, they will attempt to increase the time per order to compensate for it. The net effect of this compensation (**B6** and **B7**) is to oppose the reduction of time per order when the service center is under high work pressure.
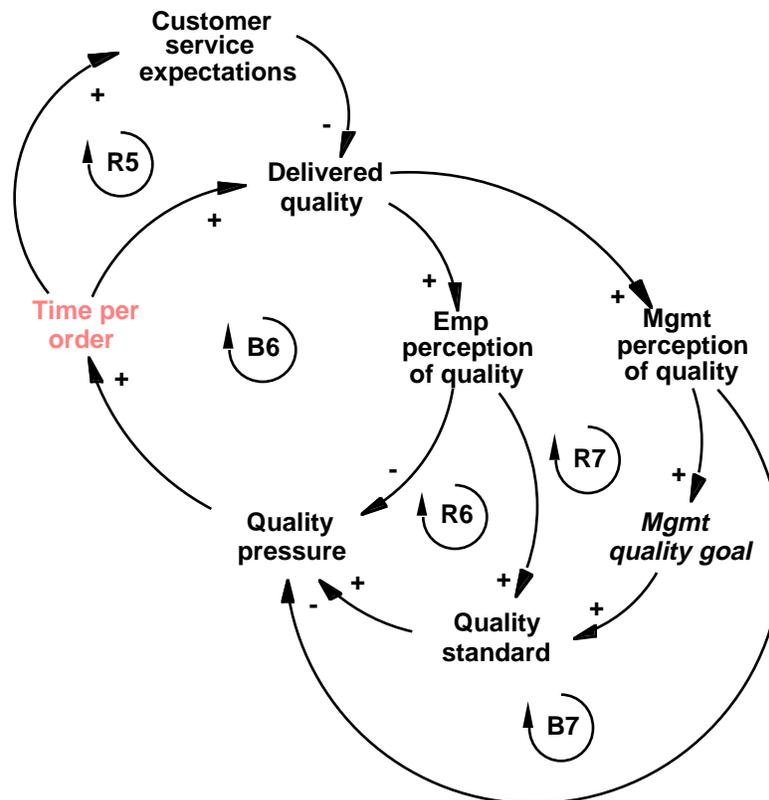
Figure 3.7   Causal loop diagram. Service quality

Nevertheless, the employees' quality standard **and** customers' service expectations are updated through an adjustment process to past performance. That is, expectations of what is appropriate service quality will be updated towards whatever the current delivered quality is. Additionally, customers' service expectations are modified by a reduction of customer heterogeneity. Low quality performance drives away the quality oriented customers leaving the firm with the price conscious customers that have lower quality standards. The expectations adjustment processes create a set of reinforcing loops (**R5~R7**) – albeit with long delays – that allows the firm to drift to low performance (Forrester, 1969; Lant, 1992; Levinthal and March, 1981).

Assuming a freeze in headcount and a commitment to deal with the service backlog, the system has a natural tendency to drift towards poor quality. If the system is perturbed, there are three balancing loops trying to bring work pressure back to normal level. The time constant for each of those loops determines their speed of adjustment. Work intensity is increased first (**B2**); if it is not capable of dealing with the pressure time per order is reduced (**B3**). If the increased order fulfillment rate does not manage to reduce the service backlog then, finally, the quality standard (**R6**, and **R7**) will erode until equilibrium is reached.

Without an explicit intervention from management to increase the quality standard, the drift to low performance becomes permanent even if the external pressures on the system are removed. If production pressure is reduced beyond its normal point, either through hiring or a reduction of orders, the slack would be absorbed by the three balancing loops in the same order that they acted to reduce high production pressure. Since the quality standard is eroded, and is the slowest to adjust, the quality pressure is diminished, thus reducing the main source of pressure to increase the time per order.

Summarizing, the dynamic hypothesis can be articulated by the following propositions (Senge and Oliva, 1993):

1. In high-contact service businesses, it is difficult to measure quality because it is intangible and subjective.

2. There is a tendency to manage service businesses by what is easily measurable. Decision makers tend to assess whether service capacity – a function of number of people, experience levels, skills, efforts, and attitudes – is adequate based on expenses and throughput figures. Management typically attempts to maximize throughput per employee and minimize expense ratios *{loops B1, B4, B5}*.

3. Because of the inseparability of service delivery from service personnel, it is relatively difficult to achieve productivity gains in a high-contact service.

4. Since service capacity determines the ability to provide services at a given quality level, and 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 *{loops B2, B3}*.

5. The consequences of reducing attention to customers are low levels of service relative to what is possible, high costs of poor quality (e.g., rework), low customer loyalty, high turnover of service personnel, and mediocre financial performance *{loops R1, R2, R3, R4}*.

6. Underinvestment in service capacity is frequently masked by eroding operating standards, so that servers *and* customers come to expect mediocre service and justify current performance based on past performance, rather than on absolute standards or goals *{loops R5, R6, R7}*.

The dynamic hypothesis can be taken to the next level: 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. Examples of industry-wide erosion of service standards – in banking, health care, insurance, and professional services among others – have frequently been cited by the popular press (Feinberg and Levenstein, 1985; Koepp, 1987; Main, 1981; Tuchman, 1980).

§3.5. Model Structure

This section contains a formal description of the service delivery theory formulated as a system dynamics model. The model consists of 55 equations – of which 18 are state variables and seven are table functions – with 51 system parameters. Efforts are made to provide evidence for the hypothesized causal relationships and theoretical foundation for the formulations. The evidence is mainly drawn from the human resources, behavioral decision theory, marketing, and operations management literature, and formulations adhere to the system dynamics approach. The theory also draws from economic theories of the service industry and TQM developments for the service sector.

Consistent with SD methodology (see Lane and Oliva, 1994 for a summary), special emphasis has been placed in describing the decision making processes depicted in the model. As an *a fortiori* assumption for the base case model, managerial decisions, i.e., hiring, capital investment, desired order fulfillment rate, and quality goal, are assumed to be made optimally and without the financial pressures identified above. Employees' operational decisions, i.e., work intensity, allocation of time per order and turnover, are formulated under more realistic assumptions concerning data availability.

§3.5.1. Boundary and Scope of the Model

The purpose of the model is to provide a theory for the erosion of quality identified in the service industry that is endogenously generated. That is, we are interested in an explanation that would permit management to intervene in the situation by modifying some of the levers that they have under their control. The model purpose defines the scope and focus of the model that are reflected in the model boundary. Table 3.3 delineates the primary features included in the model (endogenous variables), the exogenous parameters, and what is excluded from the model.

Exogenous parameters are factors that affect the performance of the service center, but that are considered to be removed from the main purpose of the model. For example, although the model tracks the technological content and labor requirements embedded in the capital acquired by the service center, the evolution of those parameters, due to technological or managerial breakthroughs, is considered beyond the scope of the model.

Endogenous variables	Exogenous parameters	Omitted in the base case model
• Service capacity • Service backlog • Work pressure • Desired time per order • Work Intensity • Fatigue • Effects of fatigue on productivity • Effects of fatigue on turnover • Service quality • Perception of service quality (by employees, management and customers) • Expectations of service quality (by employees, management and customers) • Effects of quality on turnover • Production factor demand • Production factor acquisition and discard processes • Personnel learning curve	• Customer demand • Technological evolution of capital • Price of production factors • Normal industry turnover • Professional quality standard • Competitors' quality	• Market feedback • Daily fluctuations of service capacity due to absenteeism • Financial performance of the service center • Rework because of poor service delivery

Table 3.2   Model boundary

Among the most important features omitted in the base model is market feedback, i.e., there is no explicit response from the simulated market (customer orders) to the performance of the service delivery. The exclusion of market feedback from the model implies that the service center is not facing direct competitors or that customers cannot find an alternative supplier for the service – clearly the case in the claims adjusting process described early in the chapter. Furthermore, the captive customer scenario is appropriate for customer service centers for products already purchased, internal services in an organization (information technology, personnel, maintenance, etc.), regulated industries (education), and monopolies (government services). If feedback to the market is permitted, i.e., simulated customers have a choice of service providers, the firm, in the long run reaches equilibrium by accepting only the number of customers it can service at the industry's average service quality and delivery delay[3]. The market feedback boundary

---

[3] A detailed description of the market feedback mechanisms and their behavioral implications can be found in Oliva, 1992.

was explicitly selected to focus the area of inquiry on the inner workings of the service delivery process.

Consistent with the fixed demand condition, the decisions for demand of production factors – labor and capital – are assumed as cost minimizing rather than profit maximizing. That is, customer demand is taken as given and service capacity is built up to respond to it. The model also omits any explicit modeling of rework – additional demand – generated because of poor quality. To simplify the presentation of results, a decision was made not to model explicitly the financial performance of the service center; instead, I decided to focus on the operational and quality metrics that seem to be in conflict. Translations of these metrics into financial performance should be relatively simple.

Because of the long term dynamics for the evolution of service quality, the high-frequency random fluctuations of service capacity – due to absenteeism or holidays – and customer orders have been excluded from the base model. Additionally, the long term perspective has made it possible to make certain simplifying assumptions regarding the level of aggregation within the model boundary. Specifically,

- Explicit variation in the time required to process individual orders is not represented.

- All personnel, although with different skill levels, are supposed to perform the same activities. There are no distinctions of personnel by function or responsibility (no representation of back-office as distinct from front-office operations).

- All capital resources are aggregated into a single stock and denominated in equivalent resources needed by a person to perform his or her job.

- As discussed in section 3.3, the multiple dimensions of quality have been collapsed into a single metric.

For purposes of the presentation, the model has been decomposed into three major sectors. In §3.5.1 I describe the *service capacity* sector. Capacity is formulated as a function of two production factors: capital and labor. The sector also represents the physical acquisition of the production factors, and the major determinants of their productivity. The section concludes with a description of the managerial decisions assumed for the acquisition of those resources. The next section contains the equations describing the dynamics of *service delivery*: the flow of customer orders, service backlog, and the employees' response to changes in work pressure. Finally, the evolution and

adjustment of *service quality* are detailed in §3.5.3. Figure 3.8 shows the main model subsystems, their information dependencies and the main exogenous inputs to the model.

Throughout this section, the following notation will be used for variable representation: stocks will be represented by Latin capital letters, rates and endogenous variables by Latin lower case letters and abbreviations. System parameters are represented by Greek lowercase letters. Tables 3.3 and 3.4 contain a summary of the main variables and system parameters in the model.

State variables	
A	Total technological content of capital
B	Service backlog
C	Customer expectation threshold
E	Management's perception of labor productivity
Fp	Accumulated fatigue for productivity
Ft	Accumulated fatigue for turnover
I	Total labor required by capital
K	Capital
Ks	Capital supply line
L	Total labor *
Le	Experienced personnel
Lr	Rookies
Lv	Labor Vacancies
Qc	Quality perceived by customers
Qe	Quality perceived by employees
Qm	Quality perceived by management
G	Quality goal (management)
S	Quality standard (employees)
T	Time per order

Table functions	
tw	Effect of work pressure on time per order
wi	Effect of work pressure on work intensity
tq	Effect of quality pressure on time per order
eqt	Effect of quality on turnover
efp	Effect of fatigue on productivity
eft	Effect of fatigue on turnover
q	Service quality as function of time per order

Rates	
k	Capital rates (orders, acquisitions, discards and sales)
l	Labor rates (orders, hiring, experience, turnover and layoffs)
s	Service rates (customer orders and order fulfillment)

Table 3.3   Model variables
* Total labor is not a stock in the model, however, it behaves as one (see eq. 3).
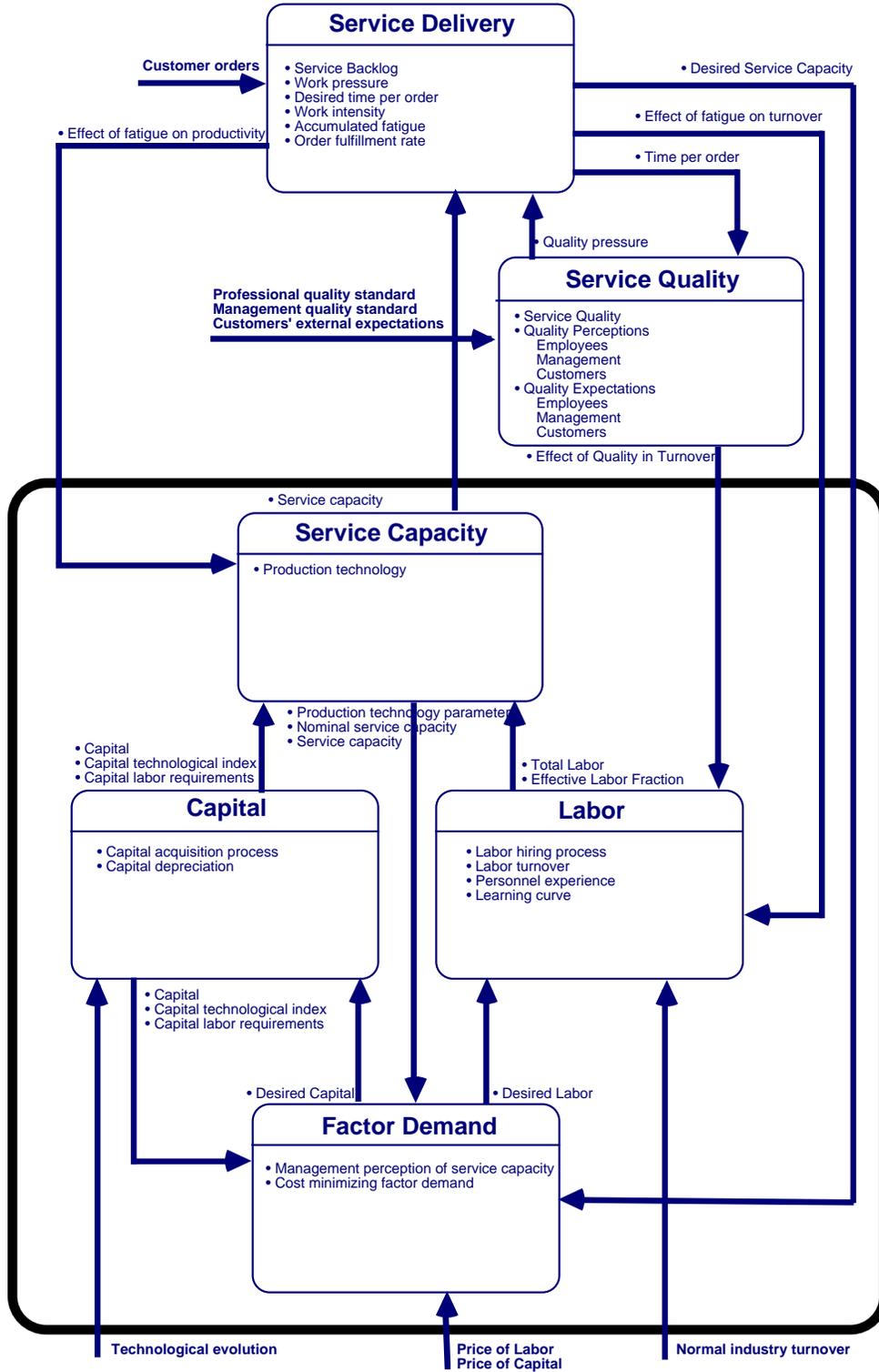
Figure 3.8   Model subsystems

Exogenous inputs are in bold.

Service Capacity	

Labor

τ_l	Time to adjust labor
λ_l	Hiring delay
τ_v	Time to cancel vacancies
τ_t	Time for turnover
τ_{ld}	Time to implement labor layoffs
χ	Maximum labor growth rate
τ_e	Time for experience
ε	Relative effectiveness of rookies
η	Fraction of experienced personnel for training

Capital

τ_k	Time to adjust capital
λ_k	Capital acquisition delay
τ_s	Time to cancel capital on order
τ_d	Average capital life
τ_{kd}	Time to implement capital sales
φ	Maximum capital growth rate
ϕ	Capital technological content
θ	Capital labor intensity

Factor Demand

τ_{pe}	Time to perceive labor effectiveness
σ	Elasticity of substitution of production factors
κ	Capital intensity
π_l	Price of Labor
π_k	Price of Capital

Service Delivery	
τ_r	Minimum residence time for an order
τ_p	Minimum processing time per order
λ_s	Desired delivery delay
τ_{to}	Time to adjust desired time per order
τ_{fp}	Time to accumulate fatigue for effect on productivity
τ_{ft}	Time to accumulate fatigue for effect on turnover

Service Quality	
α	Responsiveness coefficients for aspiration adjustment rates (8)
γ	Responsiveness coefficients for quality pressure (3)
τ_{qc}	Time to identify customers' perceptions of quality
τ_{qe}	Time for employees to adjust quality perception
τ_{qm}	Time for managers to adjust quality perception
τ_{cq}	Time for customers to adjust service expectations
τ_{sq}	Time to adjust quality standard
τ_{gq}	Time to adjust quality goal
β_c	Customers' quality perception bias
β_e	Employees' quality perception bias
β_m	Management's quality perception bias
μ	Customers' expectations reference
ψ	Professional quality standard
ξ	Management quality standard

Table 3.4   Model parameters

§3.5.2. Service Capacity

The units of service capacity are assumed to be homogeneous 'capacity hours.' A number of capacity hours are required to fulfill each customer order. The nominal capacity of the service center (sc_n) is given by a Constant Elasticity of Substitution (CES) production function of two production factors – capital (K) and labor (L) – with elasticity of substitution (σ) (Arrow, Chenery et al., 1961). Although it has been argued that there may be little possibility of substitution of production factors in a high-contact service environment, the more general CES technology was chosen over the Leontief production function (Diewert, 1971) to test the impact of technological innovation and responsiveness of the factor acquisition policies over a range of elasticities.

To facilitate the initialization and interpretation of the model, the production function has been normalized to assume that each capital unit has a required average labor intensity (*i*)

(eq. 28), and that a unit of capital, with appropriate staffing – in number, skills, and under normal work intensity – will yield the reference service capacity. The relative contribution of capital in the generation of output is given by the capital intensity parameter (κ). Capital is assumed to have an intrinsic productivity given by the average technological index (a) (eq. 26). The formulation assumes that changes in technology affect the combined productivity of capital and labor.

$$(1) \qquad sc_n = a\left[\kappa(K)^\rho + (1-\kappa)(L/i)^\rho\right]^{1/\rho} \qquad \rho = (\sigma - 1)/\sigma,\ 0 \le \delta \le 1$$

The effective service capacity (sc) is determined by adjusting the nominal service capacity by the effects of personnel experience – the effective labor fraction (e) (eq. 4) – and the effect of fatigue on productivity (efp) (eq. 38).

$$(2) \qquad sc = sc_n \cdot e \cdot efp$$

§3.5.2.1. Labor Sector

The labor sector models the acquisition, training, and turnover of the labor force. A key assumption in the model is that personnel recently hired have only a fraction (ε) of the productivity of more experienced employees and that they each reduce the productivity of an experienced person by a constant fraction (η) during their training process. The learning curve that recently hired personnel undergo (developing their skills) is modeled as an "experience chain" (Jarmain, 1963; Lyneis, 1980). Labor (L) is separated into two populations: experienced personnel (L_e) and rookies (L_r). The effective labor fraction (e), measured in full-time equivalents of experienced personnel, is given by the mix of the two populations and their relative productivity[4]. Total labor (L) and effective labor fraction (e) are used elsewhere in the model to determine the effective service capacity.

$$(3) \qquad L = L_e + L_r$$

$$(4) \qquad e = \max\left(0, \left(L_e + L_r(\varepsilon - \eta)\right)/L\right) \qquad 0 \le \varepsilon \le 1,\ 0 \le \eta$$

The stock of rookies is increased by the hiring rate (l_h), and decreased by employees becoming experienced personnel (l_e) and layoffs of rookies (l_{lr}). Experienced personnel is augmented through rookies gaining experience (l_e), and reduced by the turnover rate (l_t) and layoffs of experienced personnel (l_{le}).

---

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

$$(5) \qquad (d/dt)L_r = l_h - l_e - l_{lr}$$

$$(6) \qquad (d/dt)L_e = l_e - l_t - l_{le}$$

The experience rate (l_e) captures the transition from rookies to experienced personnel. It is assumed that rookies develop full productivity through an exponential learning curve characterized by an average training period (τ_e), and that the training period is relatively short in comparison to the average employment tenure – time for turnover (τ_t). The turnover from the experienced personnel stock is assumed exponential with an average time for turnover (τ_t). 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, thus considered exogenous to the model and captured in the average time for turnover (τ_t). However, two factors endogenous to the model are assumed to modify the average time for turnover: fatigue (*eft*) and quality (*eqt*) (see eqs. 40 and 54).

$$(7) \qquad l_e = L_r/\tau_e$$

$$(8) \qquad l_t = L_e/(\tau_t \cdot eft \cdot eqt)$$

The layoff rates are active only if total labor is greater than the desired labor (L^*). The excess labor fraction (*xl*) determines the layoff rates assuming a time to respond and implement those layoffs (τ_{ld})[5].

$$(9) \qquad xl = \max\left(0, (L - L^*)/L\right)$$

$$(10) \qquad l_{lr} = (L_r \cdot xl)/\tau_{ld}$$

$$(11) \qquad l_{le} = (L_e \cdot xl)/\tau_{ld}$$

The hiring rate depends on the firm's unfilled labor vacancies (L_v) and a hiring delay (λ_l) – the time it normally takes to fill a vacancy once it has been announced. Vacancies represent the labor orders (l_o) that have not been filled.

---

[5] The proposed formulation assumes that a proportional fraction of personnel from each of the personnel categories is to be made redundant. Should an explicit layoff policy be in place, e.g., seniority or inverse seniority, the formulations of the layoff rates could be modified to include this preference through a weight (∂).

$$l_{lr} = \partial \cdot \left((L_r \cdot xl)/\tau_{ld}\right)$$

$$l_{le} = (1 - \partial) \cdot \left((L_e \cdot xl)/\tau_{ld}\right)$$

(12)
$$l_h = L_v / \lambda_l$$

(13)
$$(d/dt)L_v = l_o - l_h$$

The indicated labor orders (l_o) are determined by three factors: i) the replacement of employees that have departed the service center, ii) the correction for any discrepancies between current labor stock and desired labor (L^*), and iii) the correction of the gap between actual and desired vacancies. The responsiveness of the policy to each of these gaps is given by the time to adjust labor (τ_l). The actual labor order rate is limited by a maximum fractional growth rate (χ) determined exogenously and, if the indicated labor orders are negative, by the number of unfilled vacancies that can be canceled. The maximum growth rate reflects the service center's absorption and training capacity. To ensure continuity of the hiring rate, the desired vacancies (L_v^*) is assumed to be proportional to the hiring delay and the current turnover rate (see (Senge, 1978; Sterman, 1989a; Sterman, 1989b) for evidence supporting this formulation).

(14)
$$l_o^* = l_t + \frac{1}{\tau_l}\left[\left(L^* - L\right) + \left(L_v^* - L_v\right)\right]$$

(15)
$$l_o = \max(-L_v/\tau_v, \min(l_o^*, L\chi))$$

(16)
$$L_v^* = l_t \cdot \lambda_l$$

§3.5.2.2. Capital Sector

The capital sector captures the firm's acquisition and disposal of capital resources. The main input to the sector is the desired capital stock (K^*). Capital is measured in generic capital units and aggregated into a single stock; a high level of aggregation considering the different kinds of capital equipment that might be necessary to deliver a service. However, since the main purpose of the sector is to capture the potential increases in productivity yielded by higher technological content in capital, the aggregation is justified. The formulations in this sector are based on a simplified version of the capital sector of the System Dynamics National Model (Forrester and Mass, 1976; Senge, 1978; Sterman, 1985a).

The net impact of capital on productivity is modeled through a set of coincident flows that track the characteristics of the acquired capital (Sterman, 1981). Specifically, the sector has a co-flow for the technological index (ϕ), i.e., the capital's productivity, and the capital's labor intensity (θ), i.e., the number of people required to staff appropriately a

unit of capital. The main outputs of this sector – capital (K), its average technological content (a), and its required labor intensity (i) – are used to estimate the nominal service capacity.

The capital stock (K) is augmented by capital acquisitions (k_a), and reduced by discards (k_d) and sales (k_s). Capital discards – the normal depreciation of obsolete capital – are assumed to follow an exponential decay process with average lifetime (τ_d). Capital sales correct for excess capital over an adjustment period.

(17) $$(d/dt)K = k_a - k_d - k_s$$

(18) $$k_d = K/\tau_d$$

(19) $$k_s = \max\left(0, \left(K - K^*\right)/\tau_{kd}\right)$$

The capital acquisition rate depends on the supply line of unfilled orders for capital (K_s) and a capital acquisition delay (λ_κ). Unfilled orders represent orders for capital (k_o) that the firm has placed but has not received.

(20) $$k_a = K_s/\lambda_k$$

(21) $$(d/dt)K_s = k_o - k_a$$

The formulation of the behavioral rule for placement of capital orders (k_o) and the desired supply pipeline (S^*) are similar to those in the labor sector (see eqs. 14, 15 and 16). The maximum rate of growth (φ) represents financial and capacity absorption constraints (Goodwin, 1951; Sterman, 1985a).

(22) $$k_o^* = k_d + \frac{1}{\tau_k}\left[\left(K^* - K\right) + \left(K_s^* - K_s\right)\right]$$

(23) $$k_o = \max(-K_s/\tau_s, \min(k_o^*, K\varphi))$$

(24) $$K_s^* = k_d \cdot \lambda_k$$

Finally, it is assumed that capital has an intrinsic technological content (ϕ) and a required labor intensity (θ), driven by factors exogenous to the model. The total technological content of the firm's capital (A) is tracked as a co-flow of capital, and the average technological content (a) is used as the technological coefficient for the production function (eq. 1).

(25) $$(d/dt)A = \phi k_a - a\left(k_d + k_s\right)$$

$$(26) \qquad\qquad a = A/K$$

This formulation assumes perfect mixing of capital such that the discarded capital has the current average technological content. Under rapid technological improvement, this formulation has a bias towards underestimating the average technological content of the firm's capital. If this assumption represents a problem, the formulation could be modified to have different vintages of capital through an aging chain (Sterman, 1981). An identical formulation is used to track the average labor intensity required by capital (*i*).

$$(27) \qquad\qquad (d/dt)I = \theta k_a - i(k_d + k_s)$$

$$(28) \qquad\qquad i = I/K$$

§3.5.2.3. Factor Demand

The factor demand sector captures management's policies for setting the desired capital stock and labor force. These policies derive the desired level for each of the production factors based on the desired service capacity (sc^*), the price of each production factor (π_k and π_l), the characteristics of the technologies used by the firm as captured in the production function, and the effective productivity of labor. The formulations assume a knowledge of the capabilities of the production technology, i.e., the exact parameters of the production function, beyond what could be normally expected from a manager in a service setting. They have been formulated as optimal strategies to show that the conditions of erosion of service quality and financial performance are generated even under such conditions (see §3.6).

The labor sector operates under the premise that although employees are using the capital resources of the firm, not all employees have the skills and/or energy required to perform the job with the same productivity. It is assumed that management perceives changes in labor effectiveness after a delay (τ_{pe}), and that their perception is adjusted through an exponential smoothing process. Labor effectiveness is determined by the ratio between the effective and nominal service capacity – the combined effects of experience (*e*) and fatigue on productivity (*efp*) (see eqs. 2, 4 and 38). The demand for production factors takes into consideration the perceived labor effectiveness (*E*).

$$(29) \qquad\qquad (d/dt)E = \left((sc/sc_n) - E\right)/\tau_{pe}$$

The formulation for labor effectiveness (E) only considers the effects of experience mix and fatigue on labor productivity; management is assumed to hire to meet demand at a normal work intensity.

The optimal demand for a production factor can be found by taking the derivative of the cost minimizing function for the CES with respect to the price of the factor (Varian, 1992)[6]. The desired capital stock (K^*) is calculated using this process and correcting for the perceived effective labor productivity.

$$(30) \qquad K^* = \frac{sc^*}{a \cdot r \cdot E} \left[\left(\frac{\pi_k}{\kappa^{1/\rho}} \right)^r + \left(\frac{\pi_l}{(1-\kappa)^{1/\rho}/i} \right)^r \right]^{(1/r)-1} \left[\frac{r\left(\pi_k / \kappa^{1/\rho}\right)^r}{\pi_k} \right] \qquad r = \rho/(\rho-1)$$

This policy represents the long-run optimal level for capital resources. A similar policy for labor procurement does not benefit from the faster response of the labor market supply. Instead, assuming a positive marginal contribution, the desired labor (L^*) is calculated at the short-run optimal directly from the production function (eq. 1).

$$(31) \qquad L^* = i \left[\frac{\left(sc^*/a \cdot E\right)^\rho - \kappa K^\rho}{(1-\kappa)} \right]^{1/\rho}$$

Although this policy for labor orders has the disadvantage that it might cause the factor mix to deviate from the long-run optimal combination, it is more responsive to changes in desired production and, when combined with the hiring policies, constitutes a reasonable short term response.

§3.5.3. Service Delivery

The service delivery sector accounts for customer orders as they flow through the service center. Service backlog (B) is formulated as a stock with an inflow for customer orders (s_o) and an outflow for the order fulfillment rate (s_f). Incoming orders are given by an exogenous market.

---

[6] The cost minimizing function for the CES production function given the desired service rate and the price for each factor is given by Varian (1992, pg. 56).

$$Cost^* = \frac{sc^*}{a} \left[\left(\frac{\pi_k}{\kappa^{1/\rho}} \right)^{\rho/(\rho-1)} + \left(\frac{\pi_l}{(1-\kappa)^{1/\rho}/i} \right)^{\rho/(\rho-1)} \right]^{(\rho-1)/\rho}$$

(32)
$$(d/dt)B = s_o - s_f$$

The desired order fulfillment rate (of^*) is formulated to achieve a delivery delay target (λ_s) based on the current service backlog. The desired order fulfillment rate is translated into the desired service capacity (sc^*) by multiplying it by the desired time per customer order (T^*).

(33)
$$of^* = B/\lambda_s$$

(34)
$$sc^* = of^* \cdot T^*$$

Changes in the desired order fulfillment rate are not necessarily reflected in the actual order fulfillment rate. A set of intermediate mechanisms mediated by the employees of the service center are assumed. The obvious response to an increase in desired production is for employees to modify their work intensity, i.e., how long they work. Work intensity can initially be increased by reducing the length of the breaks during working hours though overtime is required for larger increases of work intensity. In the model it is assumed that employees adjust work intensity as a response to work pressure (p_w). Work pressure (p_w) is defined as the normalized gap between desired service capacity and effective service capacity (Abdel-Hamid and Madnick, 1991). Work intensity (wi) in the model is normalized to the units in the production function. The employees' response to discrepancies between desired and effective service capacity is assumed to be non-linear and limited by (ω^{max}) – a physical limitation given by the hours in the week or time that the service center can be open – and (ω^{min}).

(35)
$$p_w = \left(sc^* - sc\right)/sc$$

(36)
$$wi = f\left(p_w\right) \qquad f(1) = 1, f(0) = \omega^{min}, f(\infty) = \omega^{max}, 0 \le f' \le 1$$

This formulation implies perfect perception of the effective production function, arguably, because employees have first hand experience of the effective imbalances of capacity and service demand.

Extended periods of high work intensity, however, cause fatigue that eventually undermines the productivity gains achieved through longer hours (Ehrengerg, 1971; Homer, 1985; Levin, Roberts et al., 1976; Thomas, 1993). In the model, fatigue (F_p) is captured as an exponential smoothing of the work intensity (wi) over a period (τ_{fp}). The effect of fatigue on productivity (efp) is a decreasing non-linear function.

(37)
$$(d/dt)F_p = \left(wi - F_p\right)\big/\tau_{fp}$$

(38)
$$efp = f\left(F_p\right) \qquad\qquad f\left(F_p \leq 1\right) = 1, \, f' \leq 0, \, f'' > 0$$

An alternative formulation, suggested by some models in the literature (Joseph, 1983, pg. 11; Levine, Van Sell and Rubin, 1985, pg. 490), is to separate the effects of fatigue into effects on productivity, influencing the labor effectiveness, and effects on attitudes that have a direct impact on quality. However, due to the assumption linking time per order to service quality, there is no benefit to making the distinction. Were the dimensions of service quality to be separated in multiple dimensions as suggested in §3.3, this distinction would be critical.

Additionally, extended periods of high work intensity 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 in employee turnover. The smoothing process, however, is done with a different time constant (τ_{ft}) to capture the longer time it requires for burnout to affect a quitting decision (see eq. 8).

(39)
$$(d/dt)F_t = \left(wi - F_t\right)\big/\tau_{ft}$$

(40)
$$eft = f\left(F_t\right) \qquad\qquad f\left(F_t \leq x\right) = 1, \, f\left(\infty\right) = 0, \, f' \leq 0,$$

The formulation of the decision of how much time will be allocated per order is based on a process of 'anchoring and adjustment' discussed in behavioral decision theory (Einhorn and Hogarth, 1981). In the anchoring and adjustment process, employees come to a judgment by anchoring on a pre-existing level of service (the anchor), and adjusting it to take account of currently available information (Hogarth, 1980; Tversky and Kahneman, 1974). In the case of the decision of time per order, employees are adjusting to the effects of work pressure (t_w) and quality pressure (t_q). Quality pressure is formulated similar to the work pressure, and is defined as the dissonance created by the gap between employees' quality standard and the quality performance of the service center (see eq. 55).

Because the effects of production and quality pressure are based on ratios, i.e., a given absolute difference between desired and actual performance becomes psychologically less important as actual performance increases, the adjustment process is assumed to be multiplicative (Hines, 1987; Kahneman and Tversky, 1982). The formulation reflects a hill-climbing search process that does not require knowledge of the function linking the

amount of time dedicated per customer orders to delivered quality. The search process is limited by the minimum amount of time required to process a customer order (τ_p).

$$(41) \qquad T = \max\left(t_q \cdot t_w \cdot T^*, \tau_p\right)$$

$$(42) \qquad (d/dt)T^* = \left(T - T^*\right)/\tau_{to}$$

$$(43) \qquad t_w = f(p_w) \qquad\qquad f(0) = 1, f' \leq 0$$

$$(44) \qquad t_q = f(p_q) \qquad\qquad f(0) = 1, f' \geq 0$$

Finally, the potential order fulfillment rate (of) is determined from the effective service capacity (sc) corrected by the employees' work intensity (wi) and divided by the time to fulfill a customer order (T). The order fulfillment rate is physically limited by the orders that can be processed from the backlog – restricted by the minimum residence time for an order (τ_r).

$$(45) \qquad of = sc \cdot wi/T$$

$$(46) \qquad s_f = \min(of, B/\tau_r)$$

§3.5.4. Service Quality

As discussed in §3.3, the model takes the amount of time allocated per order as a proxy to operationalize service quality. A moving threshold captures customers' expectations in terms of the time that should be allocated to each order (C). Quality (q) is derived through a non-linear function of the performance gap – the normalized difference between the time allocated per order (T) and customers' expectations of time allocated (Zeithaml, Berry and Parasuraman, 1987; Zeithaml, Parasuraman and Berry, 1990).

$$(47) \qquad q = f\left((T - C)/C\right) \qquad\qquad f(0) = 1, f' \geq 0$$

The function reflects Kano's differentiation of attributes between "must-be's" and "delighters" (Burchill, 1993; Kano, Seraku et al., 1984; Shiba, Graham and Walden, 1993). 'Must-be' requirements are those which do not lead to satisfaction when fulfilled but cause dissatisfaction when not fulfilled. 'Delighters,' on the other hand, are requirements that create satisfaction when fulfilled but do not represent dissatisfaction when not present. Figure 3.9 shows the function assumed for the base case. The shape of this relationship is also consistent with the customers' 'tolerance zone' for service quality (Parasuraman, Zeithaml and Berry, 1994; Strandvik, 1994).
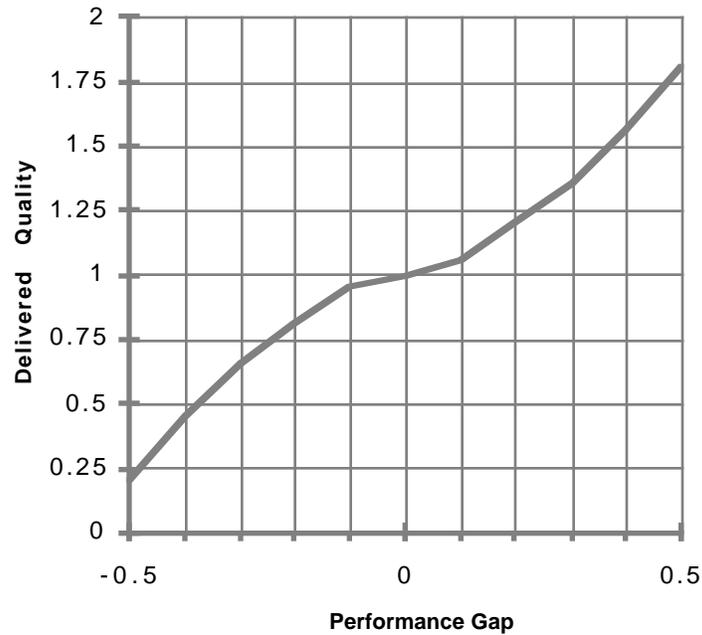
Figure 3.9   Effect of performance gap on delivered quality

Since all service quality dimensions are compressed into a single metric, this representation assumes that employees have a good understanding of the difference between 'must-be' and 'delighter' attributes, and that, if pressured by time, they would reduce the delighters before interfering with the must-be's.

Delivered quality (q) is perceived by employees (Q_e) and management (Q_m), with a bias (β), through a smoothing of the quality delivered to the customer and, if the appropriate market research instruments are in place, the customers' experience with the service delivery (Q_c) is eventually reported to management with some delays and bias.

$$(48) \qquad (d/dt)Q_e = (q + \beta_e - Q_e)/\tau_{qe}$$

$$(49) \qquad (d/dt)Q_m = (q + \beta_m - Q_m)/\tau_{qm}$$

$$(50) \qquad (d/dt)Q_c = (q + \beta_c - Q_c)/\tau_{qc}$$

The time constants for these perceptual processes are assumed to be different and ranked according to their immediacy to the delivery process. 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, the market survey instruments have an even longer detection process
($\tau_{qe} \leq \tau_{qm} \leq \tau_{qc}$).

Each of the agents involved in the service delivery process – employees, managers, and customers – in addition to the perception process described above, are assumed to have an internal goal of what level quality ought to be delivered. Customers hold an expectation (C) of what constitutes an adequate allocation of time for processing a customer order. Management has an explicit managerial goal (G) that reflects their desired level of quality. Finally, employees hold an internal quality standard (S) that can be conceptualized as the level of quality they would deliver in the absence of work pressure and under normal work intensity.

These goals are conceptualized as 'levels of aspiration' (Lant, 1992; Lewin, Dembo et al., 1944), and are adapted following a decision rule that updates aspirations based on 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.

The decision rules for adjusting these levels of aspiration are generalized to include a constant term that operates as an anchoring mechanism (Lant, 1992). If significant, the constant term has two implications for the adjustment process. First, it affects the speed at which the adjustment process updates the aspiration level, making the process more responsive to changes towards the anchor value. Second, it prevents the level of aspiration from adjusting completely to the level of performance in equilibrium.

The quality standard (Q_s) is assumed to be anchored to a fixed professional quality standard (ψ) (Levine, Van Sell and Rubin, 1985), and it adapts, with some delays, to management's desired quality goals (G) and the employee's perception of current quality delivered to the customer (Q_e).

(51.a) $$S^* = \alpha_{sp}\psi + \alpha_{sg}G + \alpha_{se}Q_e \qquad \alpha_{sp} + \alpha_{sg} + \alpha_{se} = 1$$

(51.b) $$(d/dt)S = \left(S^* - S\right)/\tau_{sq}$$

Management's quality goal (G) is anchored to an exogenous standard (ξ)[7] and it adapts to management's own perception of delivered quality (Q_m) and the feedback from customers' perception of service quality (Q_c).

(52.a) $$G^* = \alpha_{gi}\xi + \alpha_{gm}Q_m + \alpha_{gc}Q_c \qquad \alpha_{gi} + \alpha_{gm} + \alpha_{gc} = 1$$

(52.b) $$(d/dt)G = (G^* - G)/\tau_{gq}$$

Finally, customer expectations are anchored to the service provided by other suppliers in the industry (μ) (Zeithaml, Berry and Parasuraman, 1987), and adapt to the current experience of time allocated per order (T).

(53.a) $$C^* = \alpha_{ci}\mu + \alpha_{cp}T \qquad \alpha_{ci} + \alpha_{cp} = 1$$

(53.b) $$(d/dt)C = (C^* - C)/\tau_{cq}$$

An argument similar to the one presented for the time constants of the perception process for employees, managers, and customers can be made for the adjustment of aspirations. Employees, who experience every day pressure to deliver the service, adjust their aspirations of service quality relatively quickly. Next in adjustment speed are managers with direct responsibility for cost control in the service center. Finally, customers, who only experience the service delivery process every now and then, have a more inflexible level of aspiration ($\tau_{sq} \leq \tau_{gq} \leq \tau_{cq}$).

The perceptions of quality and quality goals have two major feedback mechanisms to the operations of the service center. 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 service quality (Schneider, 1991; Schneider, Parkington and Buxton, 1980; Tornow, 1991). In this model, the employees' perception of delivered quality affects the average duration of employment (see eq. 8).

(54) $$eqt = f(Q_e) \qquad f(0) = 0, f(1) = 1, f' \geq 0$$

Finally, Quality pressure (p_q) affects the time allocated per order (see eq. 44). Quality pressure is defined as an indicator of the dissonance created in employees by the gap between the quality performance of the service center and the employee's desired quality

---

[7] This standard could reflect the industry's current quality standard or the goals set by an internal quality improvement program.

level (*S*). Quality performance is a weighted average of the different perceptual mechanisms that the service center has. Although employees might believe that they are doing a good job in delivering service quality management could have a different perception ($Q_e < Q_m$) and decide to put some pressure on the employees to improve.

(55)
$$p_q = \left(S - \left(\gamma_e Q_e + \gamma_m Q_m + \gamma_c Q_c\right)\right)/S \qquad \gamma_e + \gamma_m + \gamma_c = 1$$

The next section makes some assumptions about the values of the system parameters and presents simulation results showing the range of behaviors the model is capable of generating.

§3.6. Simulation Results

In this section I will replicate, based on the structure proposed above, the reference mode described for the Hanover insurance case, and explore the range of behaviors the model is capable of generating. The different behavior modes are created by varying the strength of the alternative operational responses that management and employees might decide to enact. Senge and Sterman (1992) identified the three possible reactions to a surge of work pressure: increase the employees' work intensity, devote less time to each order or increase the effective service capacity (B2, B3, and B4 in figure 3.5). These responses will be used to examine the model's range of behaviors[8].

To make the results comparable, all simulations are performed in the 'same service setting' – the same structural and institutional parameters. The selection of parameters will be made to represent the physical and institutional characteristics of a generic service center but no reference will be made to any particular site. To avoid transient effects in the results all simulations will be initialized in equilibrium.

§3.6.1. Context for Base Case

The system parameters are grouped into six different areas. Three of these areas are considered structural constraints for the service center (conditions difficult to change): the production technology under which the service center operates, the service delivery process, and the learning curve for new personnel. The next two areas capture the institutional policies of the service center: the policies that determine the process of

---

[8] Additional behaviors are generated if the parameters that drive the perception and formation of aspiration of service quality are modified.

acquiring and managing production factors, and the way information about service quality is perceived and expectations are formed. The last area encompasses the parameters and table functions that define the behavior of the employees of the service center.

Production technology. To define a unit of capital and labor, the price of both production factors will be assumed equal ($\pi_l = \pi_k = 1$ \$/month), and the capital contribution to the process is set to 50% ($\kappa = 0.5$). The capital productivity – technological index and labor intensity – will be assumed constant throughout these simulations, and the assumption will be made that a unit of capital is required for a fully trained person to do his or her job ($\phi = 1$ hours/month/capital; $\theta = 1$ employees/capital). Some elasticity of substitution between labor and capital will be assumed ($\sigma = 0.5$) to explore the implications of technological substitution of production factors.

Service delivery process constraints. Under normal conditions and normal quality it will be assumed that each order requires one hour's effort by an experienced person to be processed. The minimum time required to process an order, even under the lowest possible quality, is assumed to be 60% of the normal time ($\tau_p = 0.6$ hours). However, because of process delays – such as waiting for investigations in the claims adjusting process – it is assumed that the fastest an order can be taken through the service center is half a month ($\tau_r = 0.5$ months). Although not a physical constraint, the desired delivery delay represents an institutional parameter that determines the responsiveness of the service center to customer orders. For the base case simulation, the desired delivered delay was set to one month ($\lambda_s = 1$ month) – twice the minimum residence time.

Learning curve. Two parameters define the shape of the employee's learning curve: the initial effectiveness that rookies have the first day of employment and the average time it takes them to become fully effective. Figure 3.10 shows the experience curve selected for the base simulation with average time to gain experience equal to 4 months and an initial productivity of 40% of an experienced person ($\tau_e = 4$ months; $\varepsilon = 0.4$). Additionally, it is assumed that every rookie requires 10% of supervision time from an experienced person ($\eta = 0.1$).

Employee effectiveness
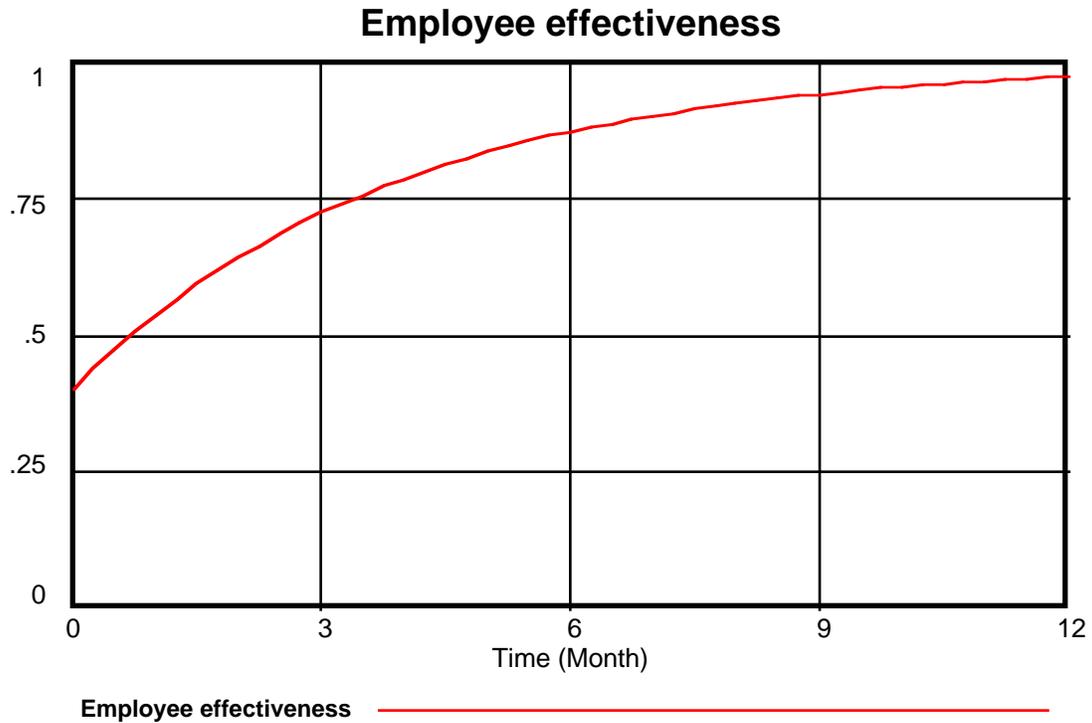
Figure 3.10   Learning curve for rookies

Acquisition of production factors. The demand for the production factors will be determined according to the cost minimizing equations described in the previous section. It is assumed, however, that management updates its perception of productivity with a time constant of six months (τ_{pe} = 6 months). The labor acquisition process – hiring and firing personnel – is assumed to be more responsive to changes in desired demand than the capital acquisition process. Table 3.5 shows the selected parameters for the two sectors.

Capital Sector	Value	Labor Sector	Value
Time to adjust capital (τ_k)	6 months	Time to adjust labor (τ_l)	3 months
Capital acquisition delay (λ_k)	18 months	Hiring delay (t to hire) (λ_l)	2 months
Time to cancel capital on order (τ_s)	1 month	Time to cancel vacancies (τ_v)	1 month
Average life of capital (τ_d)	60 months	Time for turnover (τ_t)	36 months
Time for capital sales (τ_{kd})	6 months	Time for layoffs (τ_{ld})	2 months
Max. capital growth rate (φ)	10%/month	Max. labor growth rate (χ)	10%/month

Table 3.5   Parameter values for capital and labor sectors

Service quality. The quality sector will be set to be consistent with what was observed in the Hanover Insurance case. Customers will be assumed to have a fixed service expectation equal to the initial time allocated per order ($\mu = 1$ hour), and they will not show adjustment of expectations ($\alpha_{ci} = 1$). Employees, on the other hand, will be assumed to form their standard exclusively based on past performance ($\alpha_{se} = 1$) without any pressure from management or their professional standards. To be consistent, the evaluation of quality pressure will be assumed to be based exclusively on the employees' perception of quality ($\gamma_e = 1$).

Employee parameters. Employees are continuously evaluating and responding to three different types of pressure. First, they show an immediate response to work pressure, either by increasing the work intensity (*wi*) or by reducing the time allocated per order (*tw*). Second, employees respond to the effects of accumulated work intensity (fatigue and burnout). Since the accumulation of work intensity takes time, there are some delays in the responses to this pressure. Finally, employees respond to the dissonance created by the gap between delivered quality and their internal expectations of what quality ought to be – quality pressure. Here again, the responses are delayed because of the time it takes employees to perceive changes in delivered quality and to adjust their level of aspiration.

Figure 3.11 shows the assumed responses to work pressure. Notice that the slope of the graph that describes the response on work intensity is steeper than the slope of the effect on time per order, representing a situation in which employees are more willing to work overtime than to reduce the allocation of time per order. Changing the slope of these charts in future simulations will modify the relative strength of these responses[9].

---

[9] The functions depicted in figures 3.11~3.13 hold the first or last value displayed if the input is outside the specified domain for the function.

Effect of work pressure on work intensity

Effect of work pressure on time per order
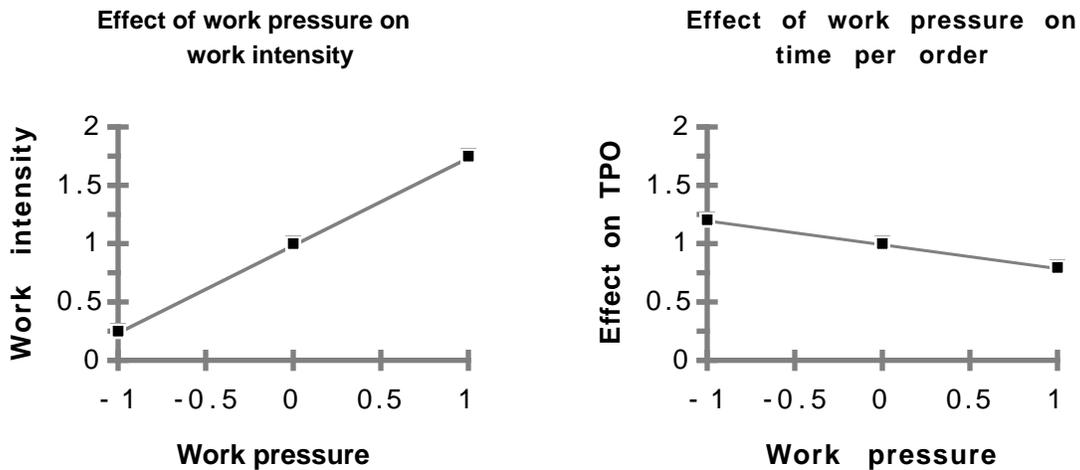
Figure 3.11   Employees' response to work pressure

Work intensity is averaged over time to determine the effects of fatigue on productivity and turnover. The average work intensity adjusts with a time constant of three weeks, (τ_{fp} = 0.75 month) to determine the effect of fatigue on productivity. The time constant for the average work intensity for the effect of fatigue on turnover is three months (τ_{ft} = 3 months). The effects of accumulated work intensity are captured in figure 3.12.

For the base case simulation it will be assumed that quality is perceived by employees without any systematic bias (β_e = 0), and that the time constant for employees to update their perception of service quality is one month (τ_{qe} = 1 month).

Effect of fatigue on productivity

Effect of fatigue on turnover
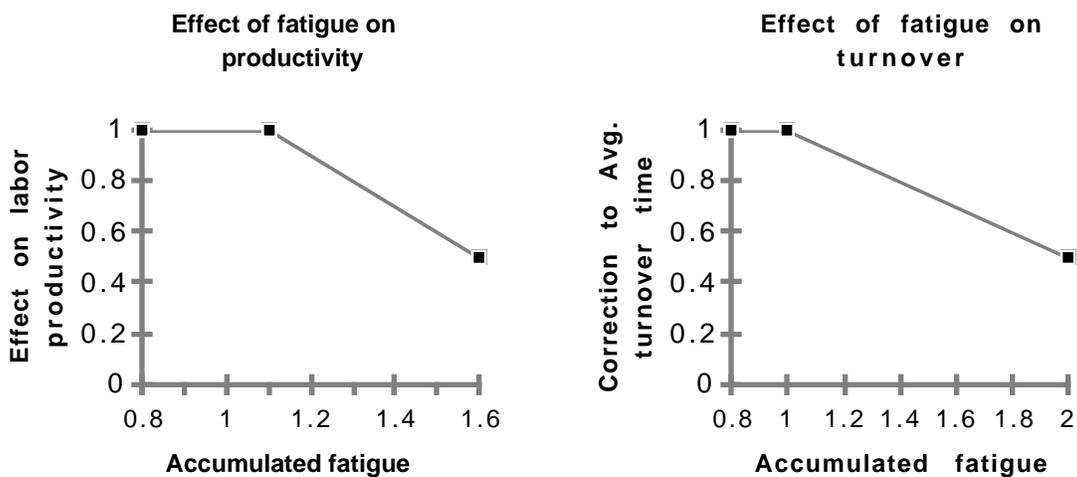
Figure 3.12   Effects of accumulated fatigue

Finally, the adjustment of the quality standard is assumed to be made with a time constant of six months (τ_{sq} = 6 months), while the time constant for the adjustment of the desired time per order is only four months (τ_{to} = 4 months). The difference in time constants ensures that in case of a reduction of time per order the slower adjustment of the quality standard will generate enough quality pressure to restore time per order. The employees' responses to the quality indicators for the base run are shown in figure 3.13.
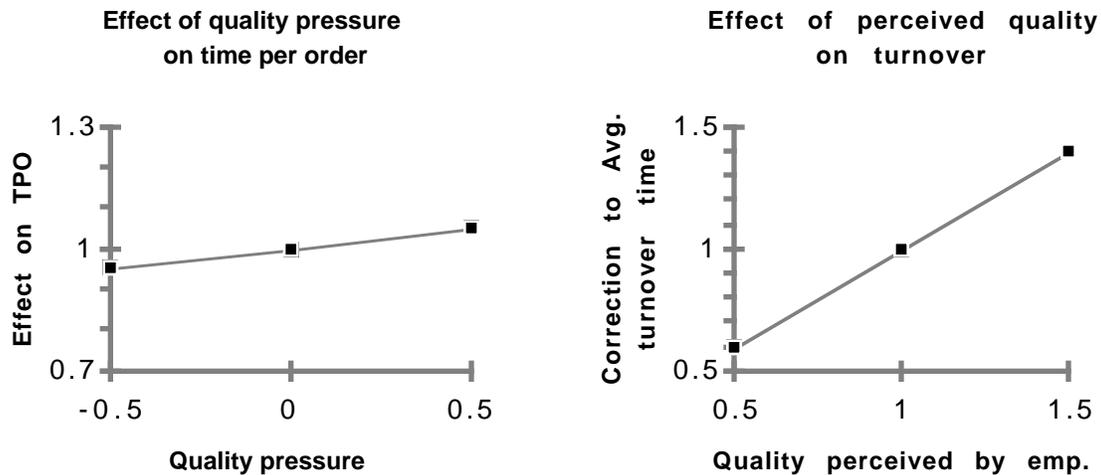
Figure 3.13   Employees' response to quality indicators

§3.6.2. Base Case

For the base case the model was initialized in equilibrium for a constant flow of one hundred customer orders per month, under the parameters described in the previous section. The model was subject to a 25% step increase in customer orders at month two. The model was simulated for three years (36 months) with a time step of 0.25 months.

Figure 3.14 shows the incoming order rate and how the desired order fulfillment rate is increased as the backlog in the service center grows[10]. The service center manages to match the order fulfillment rate to the incoming order rate by month eight, thus reducing the existing backlog and bringing down the desired order fulfillment rate. By month 12 the number of orders processed equals the desired order fulfillment rate, and the center reaches equilibrium of throughput around month 14.

---

[10] Desired order fulfillment rate follows the backlog exactly because of the selected desired delivery delay (1 month).
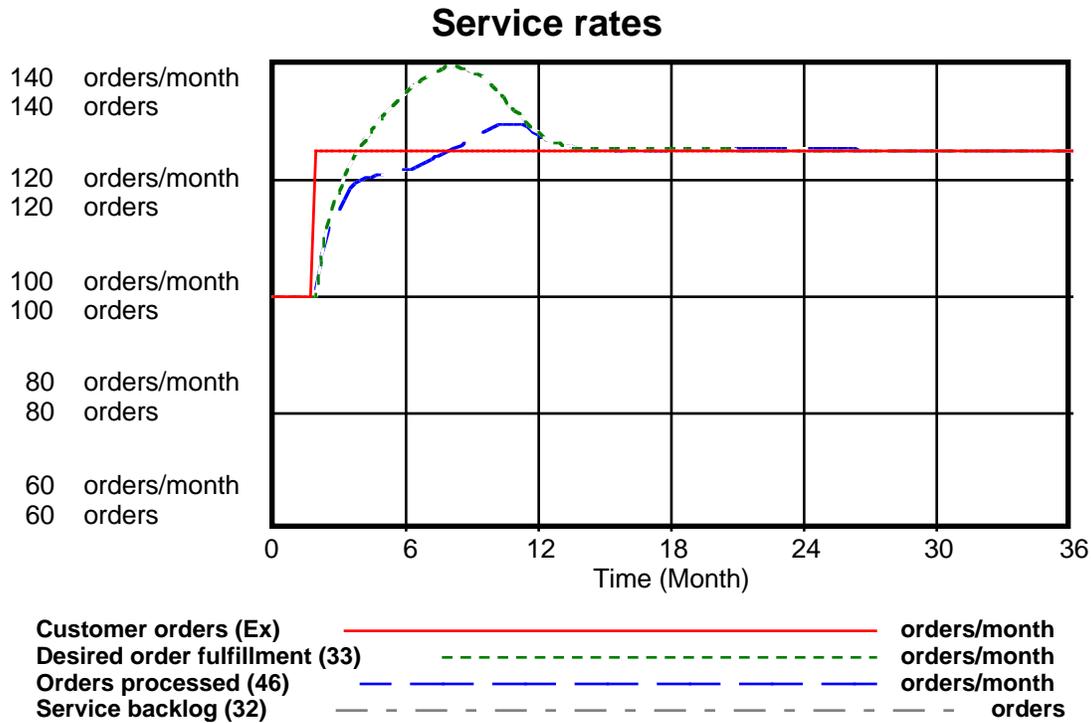
Service rates

Figure 3.14   Base case simulation results. Service rates[11]

The desired order fulfillment rate drives the desired service capacity. The increase in desired service capacity is further felt by employees as work pressure (see eqs. 34 and 35). Figure 3.15 shows the behavior over time for desired service capacity and work pressure. It is worth noting that the erosion of desired time per order causes both of these series to peak around month six, as opposed to the desired order fulfillment rate which peaks in month eight. Furthermore, by the time the system reaches equilibrium – work pressure returns to its normal operating value at month 32 – the desired service capacity has expanded only 18.6%. A 5% reduction in the desired time per order (see figure 3.17) accounts for the rest of the adjustment to match the 25% increase in customer orders[12].

---

[11] The number in parenthesis by the variable name in figures 3.14-3.19 represents the equation number in §3.5. The (**Ex**) notation represents an exogenous variable and (**S##**) denotes the definite integral from t = 0 to Time of the corresponding equation.

[12] In equilibrium the order fulfillment rate is given by $s_f = sc/T^*$. The system's adjustment for the required 25% increase in the order fulfillment rate is broken down as follows: $(1 + 0.25)s_f = (1 + 0.186)sc/(1 - 0.050)T^*$.

Initial response

140	orders/month	
140	hours/month	
.5	fraction	
100	orders/month	
100	hours/month	
0	fraction	
60	orders/month	
60	hours/month	
-.5	fraction	

Time (Month)

Customer orders (Ex)		orders/month
Desired service capacity (34)		hours/month
Work pressure (35)		fraction
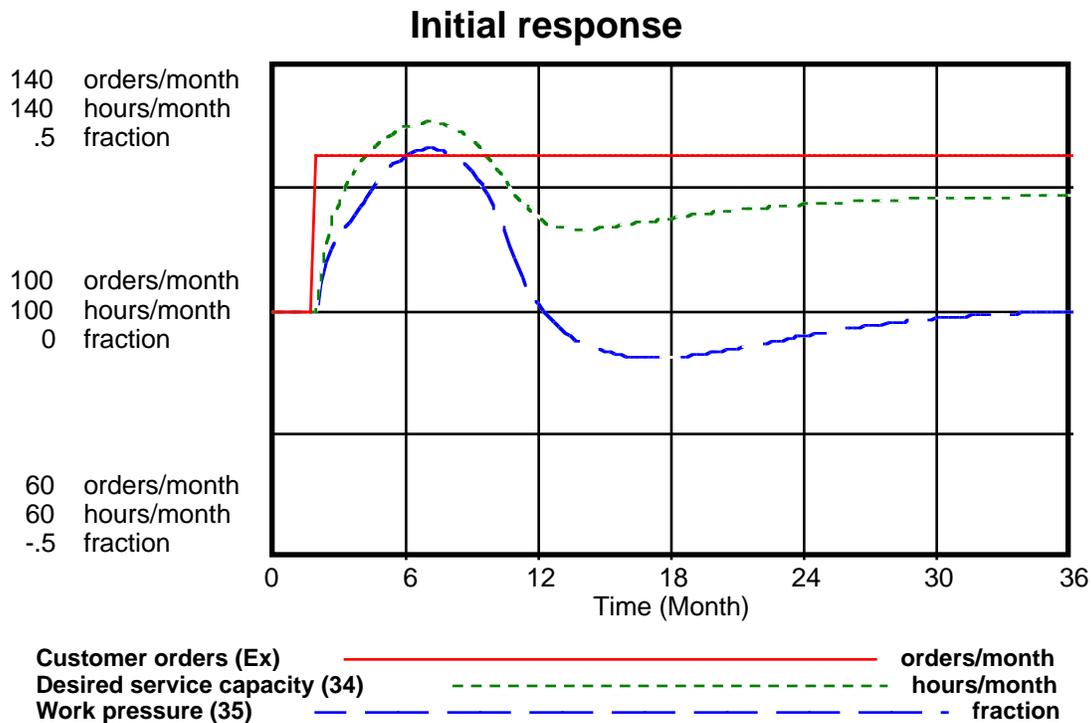
Figure 3.15   Base case simulation results. Initial response

Production factors

250	capital
250	employees
187.5	capital
187.5	employees
125	capital
125	employees
62.5	capital
62.5	employees
0	capital
0	employees

Time (Month)

Desired capital (30)	capital
Capital (S17)	capital
Desired labor (31)	employees
Total labor (3)	employees
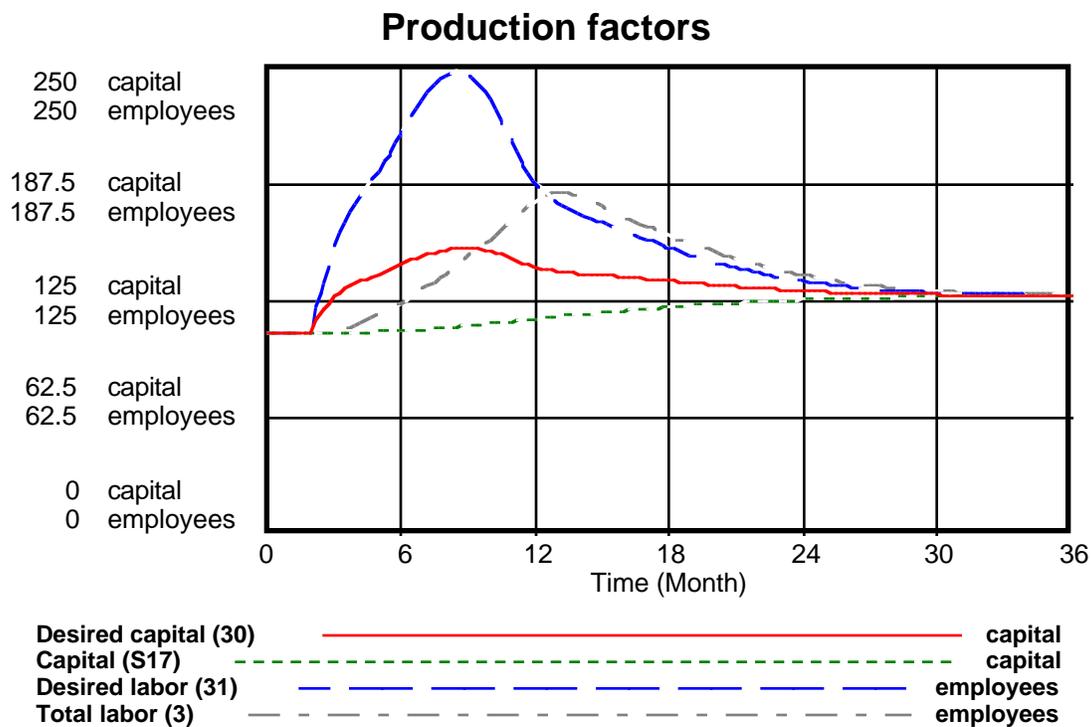
Figure 3.16   Base case simulation results. Production factors

74

Desired service capacity is used by management to determine the demand for production factors. Because of the slow acquisition process for capital, and the elasticity of substitution between production factors, management attempts to satisfy demand by increasing desired labor aggressively (see figure 3.16). Actual labor does not follow desired labor because of the constraint on the growth rate of the labor force (max. 10% per month). The desired labor force drops sharply around month eight as capital resources begin to arrive and the desired service capacity decreases. Because of the aggressive hiring total labor exceeds desired labor at month 12, bringing work pressure below its normal level (see figure 3.15). The acquisition process reaches equilibrium for both factors at month 32 – a year and a half after the equilibrium in throughput was reached.

Figure 3.17 captures the employees' response to the increase in work pressure. The first thing to notice is that these responses – the increase of work intensity and the reduction of time per order – are rapid in comparison to the capacity adjustment policies, which take over two months to generate any noticeable results. Work intensity peaks at the time work pressure peaks and drops below the normal level when the labor force exceeds the desired labor force. As the time per order drops in response to work pressure, it drags down the desired time per order. It is interesting to notice that the time per order reaches its minimum point after work intensity has been dropping for a while; a positive work pressure drives the time per order down. Time per order begins to recuperate when the effect of quality pressure compensates for the effect of work pressure on time per order.

The sustained work intensity, however, has its toll on employee productivity. Figure 3.18 shows the behavior of nominal service capacity – the theoretical capacity from the production factors available – and the actual service capacity. Actual service capacity includes the effects of skills and fatigue on productivity. During the transitional period, these two effects reduce effective service capacity even though new resources are coming into the service center, thus creating a 'worse before better' effect on the response of service capacity (Forrester, 1969). First, high sustained work intensity causes fatigue that reduces the net productivity of employees. Second, hiring of new personnel dilutes the experience mix in the service center thus reducing effective service capacity. Eventually, when the labor force matches the desired labor force (month 12), the effect of fatigue dissipates and the effective labor fraction slowly returns to its normal operating level. From figure 3.18, however, it is evident that there is a significant overshoot of nominal service capacity triggered by the lower effectiveness of actual service capacity during the acquisition stage.

Response to work pressure

1.3 fraction	
1.3 hrs/order	
1.15 fraction	
1.15 hrs/order	
1 fraction	
1 hrs/order	
.85 fraction	
.85 hrs/order	
.7 fraction	
.7 hrs/order	

Time (Month)

Work intensity (36)	———————	fraction
Time per order (41)	– – – – – – –	hrs/order
Desired time per order (S42)	— — — — —	hrs/order
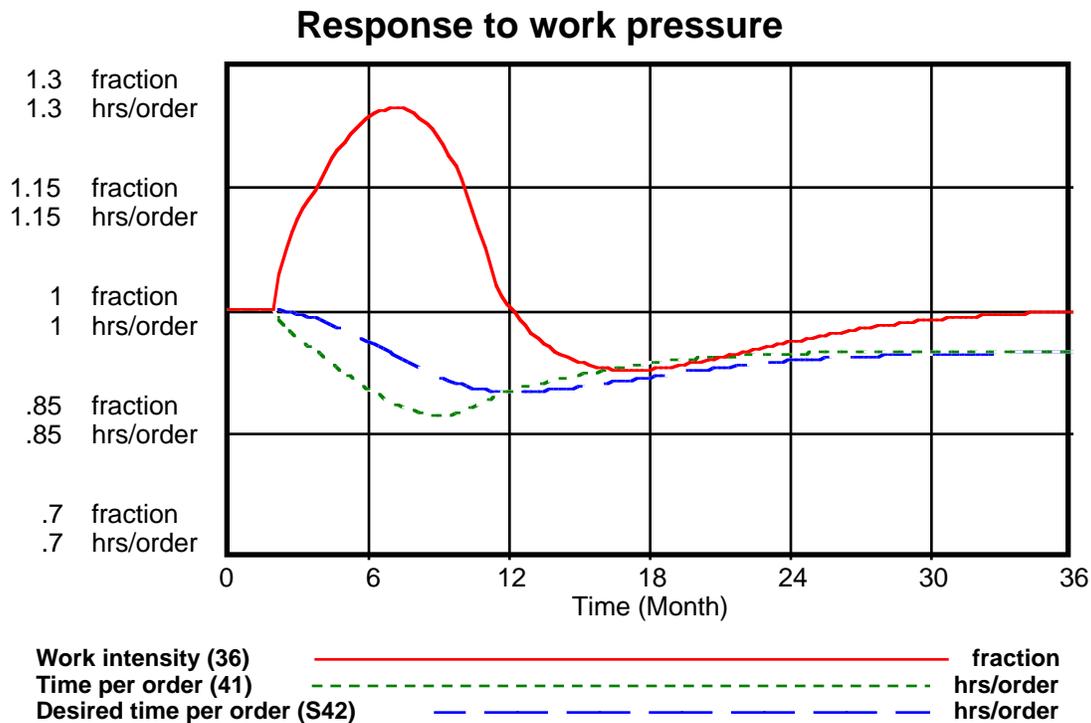
Figure 3.17   Base case simulation results. Response to work pressure

Service capacity

175 hours/month	
1 fraction	
150 hours/month	
.875 fraction	
125 hours/month	
.75 fraction	
100 hours/month	
.625 fraction	
75 hours/month	
.5 fraction	

Time (Month)

Nominal service capacity (1)	———————	hours/month
Service capacity (2)	– – – – – – –	hours/month
Effective labor fraction (4)	— — — — —	fraction
effect of fatigue on productivity (38)	— · — · — · —	fraction
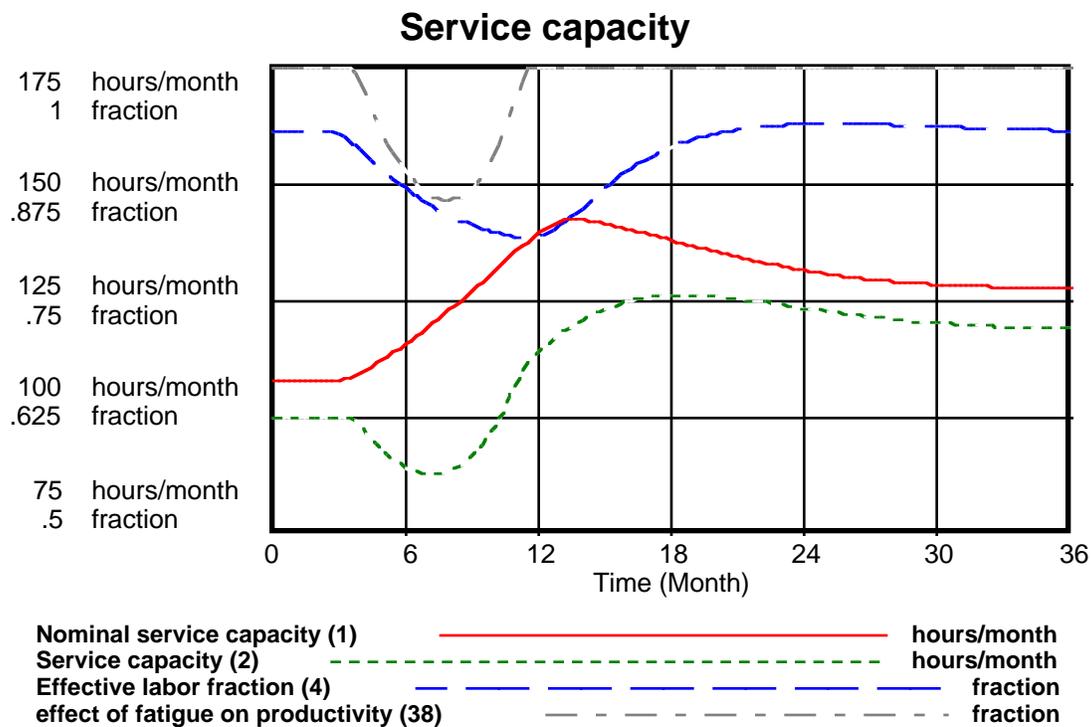
Figure 3.18   Base case simulation results. Service capacity

The excess capacity eventually helps to increase service quality. Nevertheless, quality never reaches the level it had before the increase in customer orders. Figure 3.19 shows the drop in delivered quality as the time per order is reduced. The different rates of change are due to the non-linear function that determines quality as a function of time per order (see figure 3.9). Employees' perception of service quality follows actual quality quite closely and the employees' quality standard is eroded as they become used to lower levels of service quality. Throughout this period quality pressure builds up as the gap between perceived quality and quality standard grows. When service quality finally improves in month 10 all quality indicators initiate their return to previously held values. However, because the quality standard continues to drop for a while longer than the perception of quality, quality pressure returns to its normal level faster than any other indicator, thus reducing the pressure to bring quality back to its previously held value.

In summary, the base case simulation presents a service center that experiences a 3% reduction in its quality performance as a result of a 25% increase in customer orders despite management's aggressive hiring policy to compensate for capacity shortages and a committed labor force willing to work overtime instead of reducing the time per order.
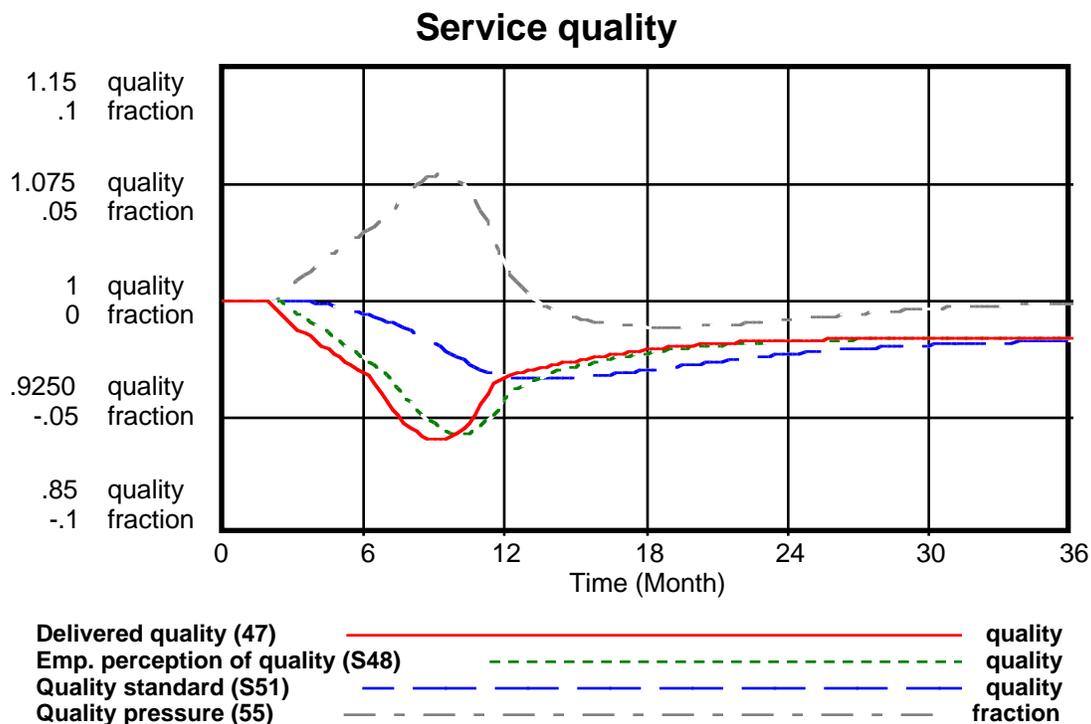
Figure 3.19   Base case simulation results. Service quality

§3.6.3. Other Simulations

This section explores the response of the model to various test inputs and the impact of structural changes in model behavior. To facilitate the analysis of results and comparison among simulations a set of summary metrics has been defined. The metrics focus on assessing the steady state of the system after a finite energy test input and the transient that the system follows to reach the new equilibrium.

Three metrics will be used to compare the equilibrium condition of various simulations. The selected metrics are: **desired time per order**, **quality standard**, and **effective labor fraction**. It is possible to capture the state of the system with these indicators because the equilibrium condition implies that all goals are being met and that there are no net forces acting to adjust any rate. All equilibrium stocks and rates in the model can be derived from these indicators, the system parameters, and the volume of customer orders.

The duration of the transient response – time to reach equilibrium – will be gauged by the time it takes work pressure to return within 1% of its neutral value. Work pressure was chosen because it is the variable that links the three sets of stocks in the system, i.e., service capacity, throughput, and quality. The extreme values during the simulation of four additional variables will be used to measure the transient behavior.

- **Max. Average delivery delay.** The average residence time of an order in the service center – from the time the order is placed to the time the order is fulfilled. This metric evaluates the center's performance in throughput and is calculated through Little's law ($d = B/s_f$) (months).

- **Max. Load per employee.** The number of orders in backlog per employee in the service center. This metric evaluates the center's adequacy of service capacity ($l = B/L$) (orders/employee).

- **Min. Delivered quality.** The quality delivered to customers. This metric evaluates the center's quality performance (see eq. 47) (quality units).

- **Min. Labor effectiveness.** The combined effects of experience and fatigue on labor productivity. This metric evaluates the degree to which labor is fulfilling its potential productivity ($f = e*efp$) (dimensionless fraction).

Table 3.6 shows the summary metrics for simulations of the base service setting described in the previous section subjected to a 10% step increase and a 25% step decrease of customer orders. The response to the 25% increase used in the base case is shown for reference.

	Change to customer orders	10% step increase	25% step increase	25% step decrease
	Final desired time per order	0.984	0.949	1.001
	Final Quality Standard	0.992	0.973	1.000
	Final Effective labor fraction	0.931	0.931	0.930
Base	Time to reach equilibrium	25	31	24
Case	Max. Avg. delivery delay	1.011	1.120	1.009
	Max. Load per employee	1.004	1.129	1.049
	Min. Delivered quality	0.984	0.910	0.998
	Min. Labor effectiveness	0.824	0.617	0.922

Table 3.6   Base case simulation results. Model response

The system response to the 10% increase in customer orders shows very little erosion of quality indicators in comparison to the base case. In looking at the transient metrics it becomes evident that the service backlog and quality indicators never departed more than 2% from their operating values. The consistency of delivery of throughput and quality is explained by an increase in work intensity. The employees of the service center were capable of dealing with the step in customer orders with overtime. Concurrently, management brought in additional service capacity fast enough to offset the effects of fatigue on productivity. Although the same policies were in place for the base case, it is obvious that the increase in workload was too much to be dealt with by work intensity exclusively. The test with the 25% decrease in customer orders was included to show how the system adjusts to excess capacity.

Three scenarios were created to explore the behavior of the system under different managerial policies and employees' responses. In the first scenario, the employees' responses to work pressure have been modified to have a more aggressive adjustment of time per order. The original 20% reduction in time per order for a unit of work pressure was changed to 40%. Simultaneously, the willingness to adjust work intensity (overtime) has been reduced from 75% to 30% per unit of work pressure (see figure 3.20). Employees under this scenario are less willing to work overtime and much more likely to reduce the time allocated to processing each order.
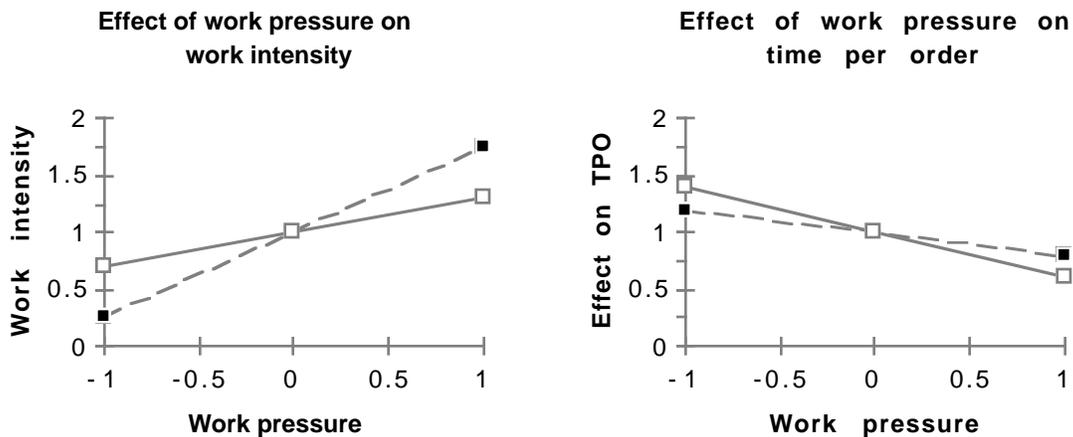
Figure 3.20   Change to employees' responses to work pressure

The second scenario increases employees' response to quality pressure to represent a setting where employees have a strong responsibility for the quality of the work delivered. The effect of quality pressure on time per order has been modify from 5% per unit of quality pressure to 40% (see figure 3.21).
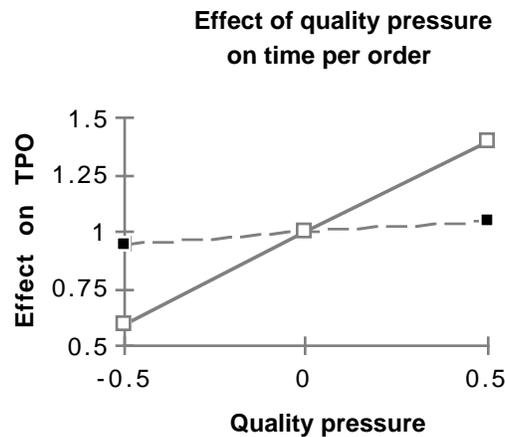
Figure 3.21   Change to employees' response of quality pressure

Finally, in the third scenario the managerial policy for acquisition of resources was modified to generate a slower adjustment process of service capacity. The change in policy was done by having desired labor to be the long term cost minimizing optimal level instead of the short term response that permitted management to hire labor aggressively while capital was being build up. Additionally, the maximum growth rates for capital and labor were set to 5%/month ($\chi = \varphi = 0.05$ %/month), and management

does not make any adjustments for loss of labor productivity in the estimation of desired service capacity (τ_{pe} = 999 months). Table 3.7 shows the indicators of the simulated response to a 25% increase to customer orders for the base case and the described policy changes.

	Policies	Base case	Aggressive WP->TPO	Aggressive QP->TPO	Slow capacity adj.
	Final desired time per order	0.949	0.921	0.949	0.874
	Final Quality Standard	0.973	0.960	0.974	0.911
25% step	Final Effective labor frac.	0.931	0.929	0.929	0.925
increase in	Time to reach equilibrium	31	26	32	17
Customer	Max. Avg. delivery delay	1.120	1.105	1.200	1.128
orders	Max. Load per employee	1.129	1.166	1.136	1.197
	Min. Delivered quality	0.910	0.878	0.926	0.886
	Min. Labor effectiveness	0.617	0.704	0.562	0.729

Table 3.7   Comparison of simulation resutls

For the two scenarios where employee responses were modified, the indicators for the final steady state are almost identical to the base case, although the scenario with the aggressive response to work pressure shows a slightly lower desired time per order and quality standard. Each scenario, however, took different paths to regain equilibrium.

In the base case, employees responded to work pressure by increasing their work intensity and attempting to maintain time per order. The effects on productivity of the extended work intensity can be seen in the minimum labor effectiveness (61%). Throughput under the base case did not suffer much – maximum 12% increase in delivery delay – and quality was maintained above 90%.

The scenario with higher concern for quality took a similar path although with much more emphasis on quality. As employees were reluctant to reduce the time per order, they had to incur more over time and average delivery delay increased – reaching a maximum 20% above desired. In the long-run, employees were not able to sustain this performance, and the effect of fatigue, combined with the introduction of new personnel, reduced labor effectiveness to 56%, bringing quality down.

The employees in the scenario with the aggressive response to work pressure did not have any problems reducing the time per order. The aggressive response brought quality down to 88% of expectations, but managed to maintain the average delivery delay within 10% of the goal. What makes this strategy interesting is that employees did not experience as

much fatigue as in the other two cases and that they were capable of dealing with the excess work load faster – work pressure returned to its normal level in month 26 instead of month 31 or 32. The reduced effect of fatigue (labor effectiveness dropped only to 70%, much of which can be explained by the introduction of new personnel), and the faster adjustment process prevented the quality standard from dropping as low as the delivered quality, thus allowing quality pressure to bring the time per customer back up.

The scenario where managerial responsiveness for the acquisition of production factors was reduced has, as expected, lower quality indicators at the end of the simulation. As service capacity takes longer to arrive, employees reduce the time per order and eventually erode the quality standard. Labor effectiveness is never as low as in the other scenarios because of the moderated introduction of new personnel into the service center. What is noteworthy is the much shorter transition time for the service center (equilibrium is reached at month 17). The reason for the shorter adjustment time is that the delayed resource acquisition policy does not overshoot the long term desired service capacity, thus work pressure is never below normal (compare to work pressure in the base case in figure 3.15). Of course, the absence of excess capacity does not permit quality pressure to bring the desired time per customer back up. The quality standard in equilibrium is very close to the minimum delivered quality.

The range of tests just presented point to some interesting characteristics of the system:

- Because of the relative delays of the various response mechanisms – changes in work intensity, time per order, and service capacity – the system has a tendency to erode service quality when facing changes in customer orders.

- If the change in customer orders is not sustained, or relatively small, employee overtime is capable of absorbing the excess demand without erosion of service quality.

- Increases of service capacity at times of high work pressure might show a 'worse before better' effect because of the time required to train new employees and the effects of fatigue on productivity.

- Given the reinforcing loops that become active after the responses from employees to high work pressure – fatigue in the case of work intensity and erosion of quality standard in the case of a reduction of time per order – the steady state of the system is not sensitive to the ranking of employees' responses.

- The steady state of the system is sensitive to the time it takes management to adjust service capacity.

§3.7. Conclusions

This chapter has presented an articulation of a theory of service delivery capable of explaining endogenously the erosion of quality in the service industry. Each assumption of the theory has been grounded in the literature and formulated into a system dynamics model. The theory was tested under a range of conditions and behavioral assumptions, and the simulated behaviors were explained in terms of the model structural assumptions, thus providing evidence for the internal consistency and construct validity of each of the relationships. Empirical validation of the overall theory will be addressed in the next chapter.

4. Empirical Validation of the Theory

§4.1. Introduction

This chapter describes a study to validate empirically the proposed theory of service delivery. A detailed argumentation of the validation methodology is first presented with an evaluation of its strengths and limitations. Next, the research site is described after a brief explanation of the selection criteria. The core of the chapter is dedicated to the calibration and adjustment of the model to the characteristics of the research site; longitudinal data, as well as semi-structured interviews and participant observation were used during the process. To assess whether we can increase our confidence in the proposed theory the chapter concludes with an evaluation of the adequacy of the model for describing the structure and behavior of the research site.

§4.2. Validation Methodology

This section, after clarifying the different criteria for model validity in the Operations Research/Management Science literature, describes the general strategy followed to test empirically the model proposed in the previous chapter.

§4.2.1. Model Validity in OR/MS

Logical empiricism (logical positivism) emerged from the Vienna Circle – group of philosophers that met during the 1920s and 1930s at the University of Vienna – and dominated science during the first half of the century. Logical empiricism regards models (theories) as canonical objective reflections of factual observations. Hence, a model is valid only if grounded in empirical observation, and if the terms used to describe it are

stated in a formal logico-mathematical language free from ambiguity. This orthodox view of science is the underpinnings of the pioneering work in Operations Research / Management Science (OR/MS) (Barlas and Carpenter, 1990; Déry, Landry and Banville, 1993). Under this perspective, the criteria put forward to judge model validity by the OR/MS community were: i) the model's representativeness of the system under study – structural and replicative validity (Churchman, Ackoff and Arnoff, 1957) – and ii) the model's predictive power (Naylor and Finger, 1967).

While being adopted as the epistemological foundation for OR, the formal and algorithmic search for the 'canonical form' embraced by logical empiricism came under attack in the second half of the century. The empirical component of logical empiricism was contested by research in sociology and psychology that showed that humans do not have access to an objective reality, and that perception is social, relational, and constructed (Cicourel, 1974; Lakoff, 1987). The philosophers of language challenged the assumption that a language – formal or otherwise – could be uniquely linked to reality (Austin, 1962; Searle, 1969).

However, the major flaw of logical empiricism was its principle of theory verification. The implication made by a theory (T) or model is that if it is true, certain data (D) would be observable (T→D). To claim that a theory (T) is verified (true) because empirical data match a predicted outcome (D) is to commit the fallacy of affirming the consequent.

> This conclusion, which derives strictly from logic, may seem troubling given how difficult it can be to make a model or develop a hypothesis that reproduces observed data. ... no matter how many confirming observations we have, any conclusion drawn from them is still an example of the fallacy of affirming the consequent. Therefore, no general empirical proposition about the natural world can ever be certain (Oreskes, Shrader-Frechette and Belitz, 1994, pg. 643).

Popper (1959) analyzed this difficulty and suggested the principle of falsification as a way to advance scientific knowledge. Under falsificationism (refutationism), empirical observations must be framed as deductive consequences of a theory – a refutable statement. If observations are shown to be true, then the theory is confirmed by those observations. If the theory fails to match observed data, then there is a certainty that the theory is untenable. The greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the model is not flawed. Falsificationism shifted the focus from theory verification to theory confirmation, and to a methodological improvement of theories through 'natural selection.'

During the 1960s and 70s, Kuhn (1970) and Lakatos (1974) showed that there are strong social and historical forces undermining theory selection under falsificationism. They provide historical evidence that scientists tend to discount data that refute their theories, and that they prefer to work with a theory that has proven to be imperfect than not to have a theory at all. Their work has led to a historical-critical perspective to explain what is valued by scientists and hence makes a scientific theory valid. Under the historical-critical perspective, theory validation becomes the process of building confidence in a theory, either through falsification or a functional perspective of theory usefulness. There is evidence that the OR/MS community is currently shifting its formal views on model validity to the more functional perspective dominant in the philosophy of science literature (Gass, 1983; Miser, 1993; Mitroff, 1972; Roy, 1993; Smith, 1993).

Within the system dynamics approach, Forrester (1961; 1968) argues that validity of a simulation model cannot be discussed without reference to a specific purpose, and has identified (1973) two groups with different interpretations of model validity emerging from their objectives for model building. He notes that most professionals (operators) take validity as relative usefulness, while academics (observers) see validity as a formal logical concept. Forrester's utilitarian interpretation of model validity is also finding its way into the OR/MS literature (Landry, Malouin and Oral, 1983).

§4.2.2. Model Calibration as Validation Strategy

The process of estimating the model parameters to obtain a match between observed and simulated distributions of a dependent variable is known as model *calibration*. Calibration of a model to an empirical setting will attest to the model's potential relevance to managers, and the generalizability of the proposed structure to other settings – external validity (Cook and Campbell, 1979). Although it is impossible to verify a model, insofar as the formulations proposed in the previous chapter are capable of capturing the behavior observed in a service setting we can augment our confidence in the theory. Since the model was specified before entering the research site, and it has been shown that it is capable of generating diverse reference modes with different parameter values, the calibration exercise constitutes a test (in the Popperian sense) for the theory.

The previous chapter presented evidence that the proposed theory is grounded in empirical observations: the theory was developed from behavior observed in a service setting – the Hanover Insurance case – and the key relationships hypothesized in the theory have been independently observed in empirical research. Furthermore, the formal

description of the model and the simulation results ensure that the theory is internally consistent and capable of providing a causal explanation of the observed behavior, thus providing a refutable causal model with multiple 'points of testing' (Bell and Bell, 1980; Bell and Senge, 1980). To test the coherence of the theory as a *whole,* it is necessary to assess whether these individual hypotheses (micro decisions) are simultaneously in place in a particular setting, and if their interactions (structure) are capable of replicating the observed behaviors of the service setting (macrobehavior).

Because the claim is being made that the model is transferable to different service settings, the focus of this chapter is the representational validity – structural and predictive – of the proposed model. The strategy for model validation is consistent with the approach proposed by van Horn: "Validation ... is the process of building an acceptable level of confidence that an inference about a simulated process is a correct or valid inference for the actual process" (van Horn, 1971, pg. 247-248). Generalizing from Naylor and Finger (1967), van Horn defines a three-stage approach for model validation:

1. Construct a set of hypotheses and postulates for the process using all available information -- observation, general knowledge, relevant theory and intuition.
2. Attempt to verify the assumptions of the model by subjecting them to empirical testing.
3. Compare the input-output transformations generated by the model to those generated by the real world (van Horn, 1971, pg. 249).

Chapter 3 described the proposed hypotheses using theory and empirical knowledge available from the literature. The present chapter describes the empirical verification of these assumptions, and formally compares the output generated by the simulation model to the behavior observed in the service setting.

The process, however, has some limitations. No model is entirely confirmed or refuted by observational data; this is particularly evident with complex hypotheses. Working in an empirical setting entails the risk that not all the data required for the calibration process are available, or that some of the hypothesized relationships are not active in the specific situation under study. Under these circumstances, the attempts to falsify the theory cannot be fully developed, i.e., no final binary decision can be made about model validity. Thus, validation is used as an inherently partial assessment of the degree of usefulness of the theory (Oreskes, Shrader-Frechette and Belitz, 1994).

Finally, the validity of the theory under the utilitarian criterion – whether managers involved in the field research believe that the theory (model) is useful and decide to implement its recommendations in their operations – was partially addressed in a report

and presentation to the managers of the research site. However, the results of the implementation, what Forrester calls the 'system improvement test,' are beyond the scope of the dissertation.

§4.2.3. Calibration Strategy

The thrust of the empirical work presented in this chapter will be the estimation of the parameters driving the behavior of the formulations proposed in Chapter 3, i.e., the model calibration process. Forrester (1961) distinguishes between two types of decisions in system dynamics models. He calls *overt* decisions those consciously made by people as part of the management or economic process, and *implicit* decisions those that arise inexorably from the current state of the system.

> ... production serves to emphasize the distinction between overt and implicit decisions. Actual, present production rate is usually the result of an implicit decision function that shows how production rate is a consequence of employment, available equipment, and materials. ... The accompanying overt managerial decisions are the decisions to attempt to hire people and to order equipment and materials (Forrester, 1961, pg. 102).

The distinction between overt and implicit decisions was used to develop a calibration strategy. Calibration of *implicit* decisions, or the parameters that drive them, is limited to identifying – through observation or interviews – the physical attributes of the workflow in the research site. Alternatively, the majority of the calibration efforts are focused on the statistical estimation of the parameters describing the model's *overt* decisions and the information processing capabilities of the agents in the service setting (Graham, 1980; Mass and Senge, 1980; Peterson, 1980; Senge, 1977). The seven non-linear functions proposed in the formal articulation of the theory are part of the overt decisions.

For each decision or set of parameters of interest, 'detailed data,' i.e., data specific to the relationship under study, were collected from the field site (see §4.3.2 Data Collection), and the parameters or shape of the relationships estimated. Three outcomes are possible from the estimation process: i) evidence found in the research site permits estimation of parameters, and confirms the hypothesized relationship and formulation, ii) evidence found in the research site leads to the rejection the proposed formulation or the hypothesized relationship, and iii) not enough data are available in the research site to test the hypothesized relationship. For the first two cases, the relationships were integrated into the model as specified by the data[1].

---

[1] Although there were a few cases where the proposed formulations had to be updated, there were no cases where the proposed relationship between variables was rejected.

In case of a lack of field data to test a micro relationship, I adhered to the system dynamics paradigm and incorporated in the model the best estimate available from the existing literature and previously available empirical research (Forrester, 1975). As shown in the previous chapter, most of these links have been tested independently in the marketing, human resource, and operations management literature, and there is some empirical evidence to support all of them.

The data gathering process was initially driven by the calibration requirements. However, when data were not directly available, or a particular formulation was not capable of generating the desired observed behavior, the process became an iterative cycle of observation, assessment, design, and model (theory) modification. To the extent these iterations were necessary, the exercise served not only as model calibration but also as theory refinement.

§4.3. The Research Site: Nelson House Lending Center

Although the proposed model is capable of generating multiple modes of behavior, I decided that a confirmation effort would be more effective if the field work was conducted at a site that showed the reference mode of erosion of service quality[2]. A visit was made in early April 1995 to explore the potential fit between the proposed research and the current business concerns of National Westminster Bank and British Telecom PLC[3]. Of the sites reviewed in the London area, I considered the NatWest's Nelson House Lending Center to be the best fit to the concerns addressed by the theory, and it was selected to perform a comprehensive field study.

In 1990 National Westminster Bank United Kingdom Branch Business – now NatWest UK Retail Banking Services (RBS) – decided to consolidate the number of branches and services under an initiative called Delivery Strategy (DS). The driving principles of DS were:

    i)   the separation of the customer interface from processing activities to enhance the quality of service provided and increase operation efficiency, and

---

[2] This does not represent selection bias since the exercise is to validate the structural mechanisms proposed as an explanation for the erosion of service quality not whether erosion of service quality actually occurs.

[3] Both organizations sponsor the "Inventing the Organizations of the 21st Century Initiative" at the Sloan School of Management.

ii) provision on expensive High Street locations, of only those activities which need to be there, removing remaining activity to more suitable, cheaper accommodation within the locality. (NatWest Bank UKBB, 1992, pg. 1).

Nelson House was created under DS as the back-office processing center for branches in the West End Region, and it currently accommodates three separate processing units: Lending Center, Service Center, and Securities (processing of legal requirements for loan collaterals).

Since the launch of DS in June 1993, the West End Region's quality metrics for the lending process – the average risk grade of the lending book[4] – have improved. However, there has been a concern in the Regional Office regarding a reduction of service quality as perceived by the customers. The origins of the erosion of service quality were believed to be the reduced in-person contact and responsiveness to specific customer needs, making the Lending Center a good candidate for a research site. From the Bank's perspective, the purpose of the research was to use the emerging theory of service delivery to explore the operations of the Nelson House Lending Center in search of evidence for and causes of erosion of service quality. The following reasons were also considered for the site selection:

a) Lending is the process with the highest cost in the bank, and significant efforts have been made to re-engineer and reduce costs in that process.

b) The process does not contain much variability, thus reducing the difficulty of calibrating the model. There is, however, evidence that indicates that finer analysis, (i.e., more time to analyze the loan request), would improve the quality of the decisions.

c) The Nelson House lending center has transactional data available since the beginning of its operations (June 1993).

§4.3.1. Site Description

The deployment plan for Delivery Strategy in the West End Region called for a 60% reduction in the number of branch outlets in the region – through consolidation of adjacent branches – and a four-year phasing in of the migration of branches into the processing centers. Figure 4.1 shows the branch absorption rate and cumulative total number of branches consolidated into the Nelson House Lending Center.

---

[4] The bank assigns a risk grade to each account based on the credit history of the of the loan holder – lower grade reflects lower risk. A Risk Index is calculated for a branch's lending book – the total outstanding loans held by that branch – by weighting the each account's risk grade by the loan amount. The Risk Index is reported monthly in the Quality of Book figures.
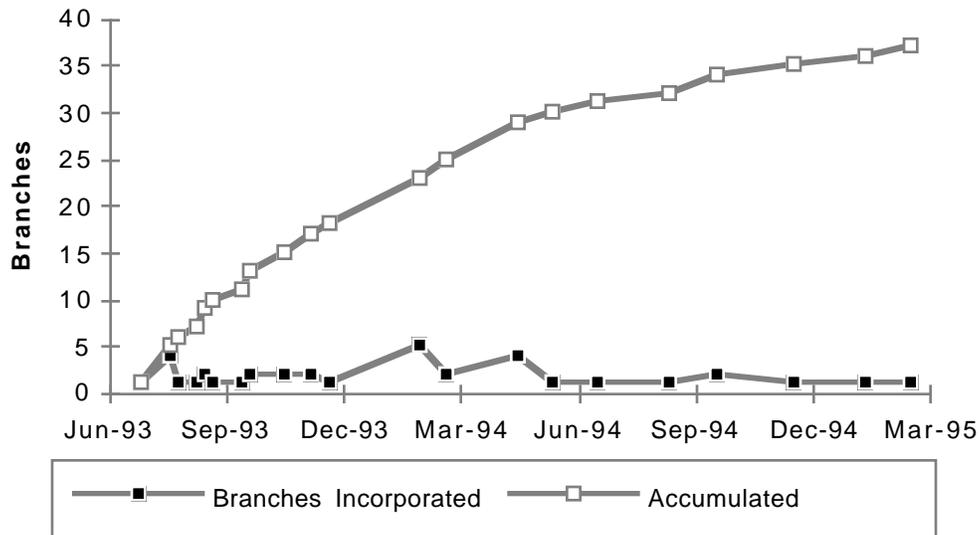
Figure 4.1   Branch absorption. Nelson House Lending Center

Since the deployment of DS, the West End Region has been merged with the City Region. As of May 1995, Nelson House was serving 245,000 accounts distributed throughout twenty branches in the West End Region – between 1% and 2% of the total account volume of UK RBS. Plans are in place to integrate five additional branches during the Fall of 1995, and six more from the City Region during 1996.

The Lending Center (LC) is responsible from making loan decisions for the mass market (personal loans and credit cards) and small business accounts (sales less than £100,000 per year) within the branches it serves[5]. Additionally, the LC controls overdraft facilities included in checking accounts. According to the LC's manager,

> The main aim of the Lending Center is to insure good quality of the lending book.
> Quality is measured by the risk grade and the amounts of unsatisfactory debt ... we also
> have to look into the healthy maintenance of that portfolio.[6]

The LC work is highly automated and performed by workgroups connected through a PC LAN with access to the bank's main databases and applications for the processing of loan applications. Workgroups are typically formed by six loan officers supervised by an assistant manager, and, although the group functions as a unit, there is a grading of responsibilities within the group members. The standard work week for a lending officer

---

[5] Relationship Centers were created in parallel to the Lending Centers to serve large corporate accounts.

[6] DR 6/27/95

is 35 hours, with the option of paid overtime if necessary. Work arrives at the LC from three sources:

a) <u>Phone</u>. Direct customer inquiries are first handled by Telephone Liaison Officers (TLO's) – housed in another facility – or directly at the branches and, if necessary, forwarded to the staff in Nelson House.

b) <u>Mail</u>. Direct customer requests and communications with the branches. Mail arrives twice a day, and is processed throughout the day.

c) <u>Computer</u> <u>generated</u> <u>reports</u>. Information technology is used to monitor the account portfolio. Daily reports identify problematic accounts that require immediate action from the LC – overdrafts, missing payments, etc. Reports that monitor the risk grade and credit limits for the accounts are generated every quarter.

The output for most processes triggered by these requests is either a letter or a phone conversation with the customer. The variety of tasks to be performed is limited, and there is a standard throughput measurement for each task. Employees are asked to maintain a tally of tasks performed during the day, and report them at the end of the week – the form to keep the tally has a classification for 42 different tasks. These reports are consolidated into the Branch Operating Report, which is submitted once a month to the Regional Office.

§4.3.2. Data Collection

The data requirements for the calibration process can be broadly grouped into four areas:

       i) Operational performance of service delivery
- Work flows
- Service Backlog

       ii) Determinants of Service Capacity
- Personnel
  - Learning Curve
  - Turnover rate
- Capital
  - Technological content of capital
  - Capital staffing requirements

       iii) Indicators of Service Quality

       iv) Managerial Policies
- Hiring
- Investment
- Service throughput goals
- Service quality goals

The most useful data to capture the relationships I am interested in testing are institutionalized norms of the service center, frequency of events over time, and time series for operational metrics. From the institutional norms it is possible to articulate the recurring decision making policies, while frequency distributions and time series data can be used to estimate some of the parameters describing the relationship between variables. Zelditch (1970) suggests interviewing informants as the best method to capture institutionalized norms and incidents about the service delivery process, while enumerating samples and archival records are the best methods to capture frequency distributions.

To be able to separate 'measurement variance' and increase convergent and discriminant validity it is desirable to have more than one method of construct measurement (Campbell and Fiske, 1959). Ideally, the measurement methods should be as different as possible so that the weakness of any method can be counterbalanced by other methods with different weaknesses (Webb, Campbell et al., 1966). Triangulation of methods can be achieved to the degree that different methods can be used to measure the same relationship (Denzin, 1970).

In addition to collecting the time series for the operational metrics of the service center – employment, work flows, and service quality – I participated in over 12 hours of direct observation of the service delivery process, and conducted 27 interviews with employees responsible for the delivery process, their managers, and people from the support organizations in the Regional Office – personnel, sales, etc. For the 15 interviews with the service delivery personnel, I followed semi-structured protocols (Cannell and Kahn, 1968; Thomas, Unpublished ms.). Given the nature of the research, the interviewees were not required to stay within the standard questions. If a profitable avenue was being pursued, the interviewee was allowed to continue in that direction. All interviews were taped and the quotes are taken from the transcriptions of those tapes. Several interviewees were subsequently contacted to clarify issues or elaborate on particular points.

§4.4. Empirical Calibration

The calibration process is presented in the same order as the sectors of the model were presented in the preceding chapter. For each sector I first provide an overview of the assumptions made by the model structure based on the its underlying differential equations. Next, the parameters that need to be estimated are identified, and a description of the quantitative data from the service center used for the estimation are presented.

Finally the estimation process is described – either from interview or longitudinal data – and an interpretation and analysis of the results is provided. Analysis of the major findings and evaluation of the overall model behavior are presented in the next section.

§4.4.1. Service Capacity

For the case of the lending center both labor and capital are required to process a customer order – specifically, each loan officer must have a computer, without which it is impossible for the loan officer to access account files. There can therefore be no elasticity of substitution of production factors, hence $\sigma = 0$. The CES production function is not defined for σ=0, but it approaches the Leontief technology in the limit $\sigma \rightarrow 0$ (Varian, 1992). The production technology in the lending center is best represented by a Leontief production function where the amount of output is limited by the resource with minimum availability. The nominal service capacity (see eq. 1 in chapter 3) will be replaced by the following formulation.

$$sc_n = a \cdot \min(L/i, K)$$

The next subsections describe the parameter estimation for the production factors.

§4.4.1.1. Labor Sector

A) Sector structure

The labor sector models the acquisition, training, and turnover of the labor force. The main inputs to this sector are the desired labor (L^*) and the effects of fatigue (*eft*) and quality on turnover (*eqt*). The labor force is depicted by two stocks – rookies and experienced personnel – that differentiate employees' productivity as they progress through a learning curve (Jarmain, 1963; Lyneis, 1980). The amount of labor is controlled by a stock-management system as proposed by Sterman (1989b). The effectiveness of the labor force is determined by the mix of employees in the two stocks and their relative productivity (see eqs. 3–16 in §3.5). Figure 4.2 shows the stock and flow structure for this sector.
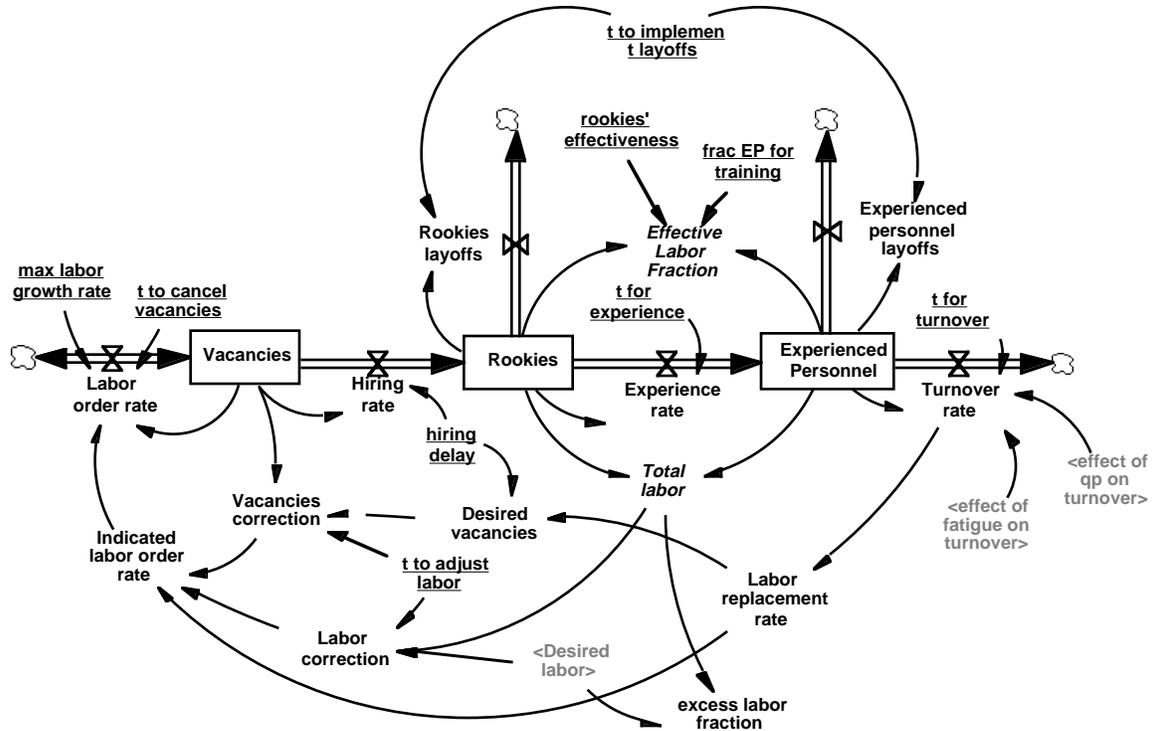
Figure 4.2   Structure of labor sector

The following parameters need to be estimated for this sector:

- Time to adjust labor (τ_l)
- Hiring delay (λ_l)
- Time to cancel vacancies (τ_v)
- Time for turnover (τ_t)
- Time to implement labor layoffs (τ_{ld})
- Time for experience (τ_e)
- Maximum labor growth rate (χ)
- Relative effectiveness of rookies (ε)
- Fraction of experienced personnel for training (η)

B) Data available

To calibrate the main parameters for the labor sector three data series provided by the service center were used.

Authorized Labor (AL): Number of employees authorized to work in the Lending Center. The time series for the first year of the center's operations (June '93-June '94) was obtained from a study done by the center's

manager. The Branch Operating Reports provided monthly data for the June '94 to May '95 period.

Total Labor (TL): Total number of employees working in the Lending Center. The series was directly available from the monthly payroll for the first year of the center's operations (June '93-June '94) and from the weekly Branch Operating Reports for the June '94 to May '95 period.

Hiring Rate (HR): Number of employees hired per unit of time. Calculated from the total personnel data series and the turnover data. Turnover data were obtained from the payroll and a report from the assistant manager of the center. Available on a monthly basis for the first year of operations and weekly from June '94 to May '95.

Figure 4.3 presents the data series for authorized personnel, total personnel, and the hiring rate.
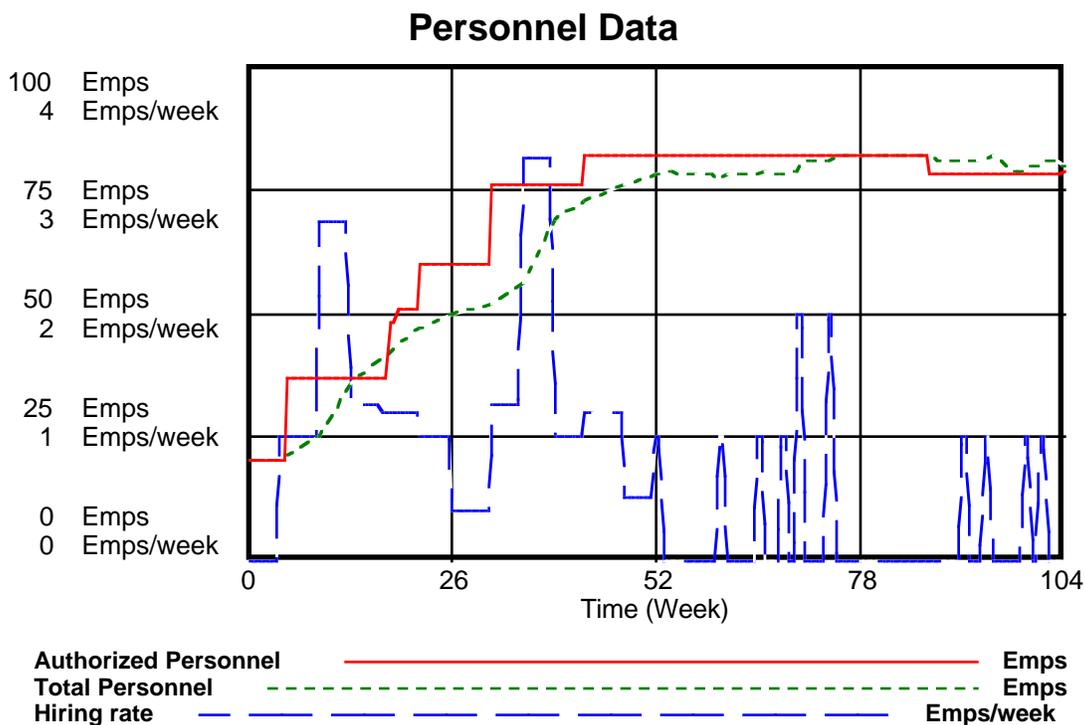
Figure 4.3   Personnel (historical data series)

C) Parameter estimation

For calibration purposes, the sector was initialized to the conditions at the opening of the center's operations. The Rookies stock (L_r) was initialized to twenty, while Vacancies (L_v) and experienced personnel (L_e) were set to zero.

`Time to implement labor layoffs and cancel vacancies.` The data available for the labor adjustment process do not show a significant decline in authorized labor that would permit the estimation of the parameters regulating the downsizing process in the lending center. Although one of the goals of Delivery Strategy was to reduce the size of the labor force in the West End Region, the downsizing efforts were focused on the closing of branches and the reduction of staff in the remaining branches. The lending center experienced a continuous expansion throughout the deployment phase of DS. Even in the first quarter of 1995, when the total number of employees in the lending center was above the authorized personnel level (see figure 4.3), personnel adjustment was happening through natural turnover, and, consistent with bank tradition, no layoffs were ever implemented. For calibration purposes, the two rates that reflect the downsizing processes in the model were made inoperative. The time to cancel vacancies was set up, based on conversations with the manager of the lending center, to one week.

`Time for experience and turnover.` The data available do not distinguish between experienced personnel and rookies. Although it is theoretically possible to estimate the time to gain experience (τ_e) from the observed performance of the workforce, this would assume knowledge of the relative effectiveness of rookies and the fraction of time dedicated by experienced personnel to train them. Neither of these parameters are directly obtainable.

The interviewees' recollection of the time required to become fully effective were diverse. The stated range for the training period was from "fully effective from day one"[7] to requiring "several months to a full year."[8] The median and mode for the self-reported time to become fully effective was three months, thus time for experience (τ_e) was fixed in the model at 12 weeks.

With time for experience (τ_e) in place, the hiring rate (HR) and the equations that link it to total labor (TL) were used to estimate the time for turnover (τ_t) through non-linear least

---

[7] SW 6/30/95
[8] JB 6/29/95

squares estimation using Powell's (1969; 1972) optimization algorithm as implemented in Vensim® (Ventana Systems, 1995)[9].

$$Min_{\tau_t} \sum_{t=1}^{104} \left(L(\tilde{t}) - TL(\tilde{t}) \right)^2 \quad \text{for } \{\tilde{t} \mid \tilde{t} \in t \text{ and } TL(\tilde{t}) = value\}$$

Subject to

$$L(t) = L_r(t) + L_e(t)$$

$$L_r(t) = L_r(t - dt) + dt\left(HR(t) - L_r(t - dt) / \tau_e\right)$$

$$L_e(t) = L_e(t - dt) + dt\left(L_r(t - dt) / \tau_e - L_e(t - dt)/\tau_t\right)$$

$$L_e(0) = 0 \qquad\qquad L_r(0) = 20$$

$$\tau_e = 12$$

The estimated value for time for turnover (τ_t), with a 95% confidence interval, is shown in the following table[10]:

```
392.165 <= T FOR TURNOVER = 400.784 <= 409.798
```

The fit between the simulated series and the historical data is presented in the following table[11]:

```
Summary Statistics for Historical Fit
Total Labor (from Hiring rate)
----------------------------------
n =   65.0

R^2                          0.997
Mean Abs. Percent Error      0.009
Mean Square Error            0.687
Root Mean Square Error       0.829
Bias                         0.004
Variation                    0.008
Covariation                  0.986
----------------------------------
```

The bias, variation and covariation are the Theil Inequality Statistics describing the fraction of the mean square error between simulated and actual series that is due to unequal means, unequal variances and imperfect correlation, respectively (Theil, 1966). Low bias and variation fractions indicate that the error is unsystematic (Sterman, 1984).

---

[9] When hiring rate data was available only on a monthly basis the hiring rate was assumed uniform for the intermediate weeks.

[10] The 95% confidence interval for the parameters is calculated from the curvature of the response surface without assumptions of symmetry. The resulting values were not sensitive to changes in dt for dt<1.0 weeks.

[11] A custom Vensim module was built after Sterman (1984) to calculate the summary statistics for the fit between the simulated and historical time series (Oliva, 1995).

Although it is surprising at first to find an average job tenure of over seven and a half years, no evidence was found from interviews or the longitudinal data suggesting that work intensity or service quality were in the range to have an impact on the turnover rate. Under such extreme conditions, the effects of fatigue and quality on turnover (*eft* and *eqt*) are undetectable. The low turnover rate is consistent with comments obtained from the interviews and the high unemployment rate prevailing in the Greater London area – above 10.5% since 1992 (Central Statistical Office, 1995). Employees' explanations included:

> The majority of people that are here are by choice. I would not expect high turnover.[12]

> Senior Management decides whether you can go or not ... [during the first year] they would not release anyone for transfer.[13]

> [Turnover is] surprisingly low. I'm sure it would be different if the economy was not what it is.[14]

Sensitivity analysis was performed on the value of time for experience (τ_e) – values between one week and one year – to determine its impact on the confidence interval for time for turnover (τ_t), and the fit to the historical data series. Variations in time for experience had no significant impact on the fit of total personnel.

`Maximum labor growth rate, time to adjust labor and hiring delay.` The original formulation for the maximum labor growth rate (χ) was built to limit the response of an established service center to changes in demand. The data available for the Nelson House Lending Center capture the initial buildup of the center, thus yielding unrealistic sustainable growth rates. The data show a maximum instantaneous growth rate (hiring rate/total personnel) of 11%/week in week nine, and it shows a steady decline as the center reaches its pre-determined size. Because of the smoother responses that the formulations in the model generate, the maximum labor growth rate for the model was fixed at 6%/week (almost 2,000%/year). The maximum growth rate was never reached in the simulation with the final parameters, effectively disappearing as a constraint.

---

[12] RP 6/30/95
[13] CR 6/30/95
[14] TV 6/28/95

With the maximum labor growth rate (χ) and the time to cancel vacancies (τ_v) in place, the authorized labor (AL) and total labor (TL) data series were used to estimate the time to adjust labor (τ_l) and the hiring delay (λ_l) through non-linear least squares estimation[15].

$$\underset{\lambda_l, \tau_l}{Min} \sum_{t=1}^{104} \left(L(\tilde{t}) - TL(\tilde{t}) \right)^2 \qquad \text{for } \left\{ \tilde{t} \mid \tilde{t} \in t \text{ and } TL(\tilde{t}) = value \right\}$$

Subject to

$$L(t) = L_e(t) + L_r(t)$$

$$L_e(t) = L_e(t - dt) + dt\left(L_r(t - dt)/\tau_e - L_e(t - dt)/\tau_t \right)$$

$$L_r(t) = L_r(t - dt) + dt\left(l_h(t) - L_r(t - dt)/\tau_e \right)$$

$$l_h(t) = L_v(t - dt)/\lambda_l$$

$$L_v(t) = L_v(t - dt) + dt\left(l_o(t) - l_h(t) \right)$$

$$l_o(t) = \max\left(-L_v(t - dt)/\tau_v, \min\left(l_o^*, L(t - dt) \cdot \chi \right) \right)$$

$$l_o^*(t) = l_t(t) + 1/\tau_l \left[(AL(t) - L(t - dt)) + \left(L_v^*(t) - L_v(t - dt) \right) \right]$$

$$L_v^*(t) = \lambda_l l_t(t)$$

$$L_e(0) = 0 \qquad L_r(0) = 20 \qquad \tau_e = 12 \qquad \tau_v = 1$$

$$\tau_t = 401 \qquad \chi = 0.06 \qquad \lambda_l, \tau_l \geq 1.0$$

The following tables show the results of the estimation process with a 95% confidence interval, and the summary statistics of the historical fit of the sector driven by the authorized labor series (AL) to the total labor series (TL).

```
* Means a bound was reached
7.6407 <= HIRING DELAY =      7.9809 <= 8.3619
1*      <= T TO ADJUST LABOR = 1       <= 1.0925


Summary Statistics for Historical Fit
Total Labor (from Desired Labor)
---------------------------------
n =   65.0

R^2                          0.982
Mean Abs. Percent Error      0.024
Mean Square Error            5.690
Root Mean Square Error       2.385
Bias                         0.009
Variation                    0.203
Covariation                  0.788
---------------------------------
```

---

[15] The $\lambda_l, \tau_l \geq 1.0$ constraint limits the values of the estimate to values that can be determined by the data. It is not possible to estimate the value of time constants shorter than the frequency of the data.

The estimate for hiring delay (λ_l) has a reasonable value of two months, while the time to adjust labor (τ_l) is on its lower bound. Both estimates have tight confidence intervals.

The rapid time to adjust labor (τ_l) is consistent with the large increments of authorized personnel observed in the data series. From an interview with the Manager of the service center it became evident that there was a well-defined staffing plan.

> Originally, when it was projected about two and a half years ago, it [staffing] was based on a head count on the number of accounts projected to come into the center ... Network [planning office] decided on a figure of, say, one member of staff for every 600 business accounts. As new branches were incorporated into the lending center the authorized personnel was updated.[16]

The fact that total labor (TL) exceeds authorized labor (AL) for four weeks around week 16 (see figure 4.4), is evidence that the service center was acting on the staffing plan before additional labor was authorized. The proposed formulations are not capable of dealing with this foresight of information. The rapid updating of the labor order rate is reflected by the estimate for the time to adjust labor (τ_l) being on its the lower bound.

Total Labor

Athorized Labor	Emps
Total Labor historical	Emps
Total Labor simulates	Emps
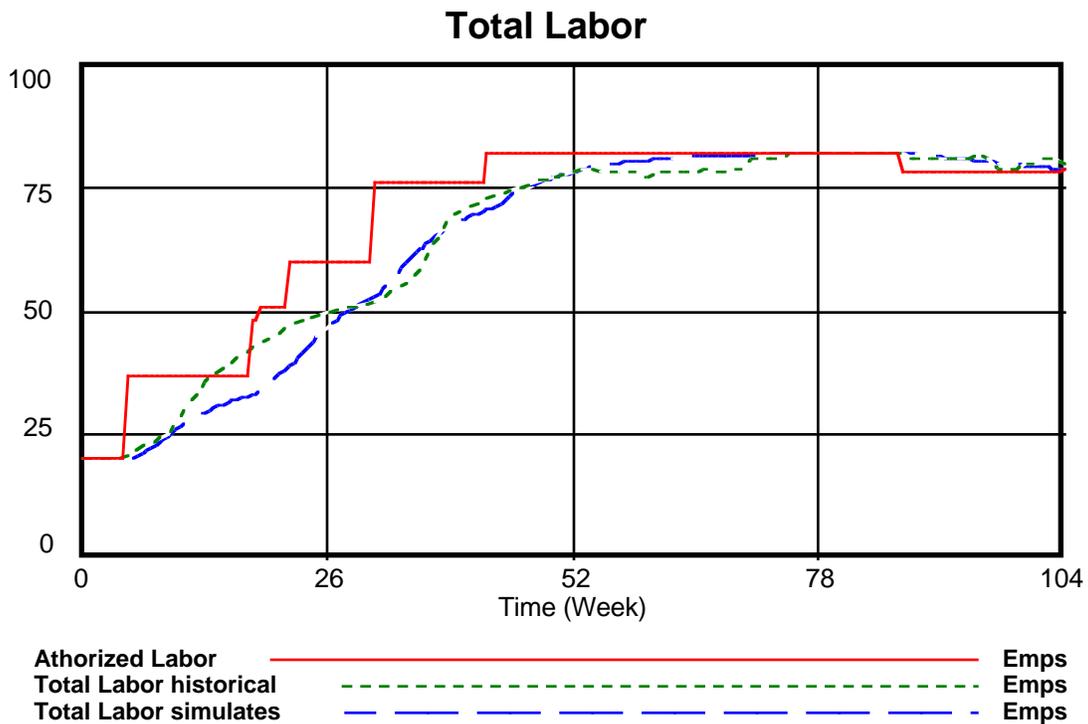
Figure 4.4   Total labor (historical and simulated data series)

---

[16] DR 6/27/95

The fit is excellent overall. About 20% of the MSE is due to unequal variance. The cause appears to be the use of anticipated information about the authorized labor series in the hiring process. Since foresight information was used only once in the growth phase of the service center, and total labor is not expected to show cyclical patterns, the error can be considered unsystematic (Sterman, 1984).

According to the manager of the lending center, the proactive hiring policy was dropped after the first year of operations of the lending center, i.e., once the absorption of new branches was concluded. To test this statement and have a better sense of the hiring policies in the equilibrium conditions, a new estimation of the time to adjust labor (τ_l) and the hiring delay (λ_l) were made with the data from the second year of operations (week 52 to 104).

```
7.4666 <= HIRING DELAY =      8.9847 <= 10.6578
7.3955 <= T TO ADJUST LABOR = 8.8996 <= 10.5691
```

The results of the estimation confirm the change in policy suggested by the manager of the lending center. Although the time to complete the hiring process – hiring delay – stayed at the same level (the two estimates are not statistically different), the time for management to adjust labor took a more realistic value of a little over two months. These results are consistent with the fact that the hiring process takes place in a centralized office that serves the whole UK Retail Banking Service, thus, should not be affected by the changes in one of many operating centers. The following table shows the historical fit to the total labor series for the second year.

```
Summary Statistics for Historical Fit
Total Labor (from Desired Labor) Week 52 to 104
----------------------------------
n =  53.0

R^2                            0.725
Mean Abs. Percent Error        0.009
Mean Square Error              0.769
Root Mean Square Error         0.877
Bias                           0.001
Variation                      0.280
Covariation                    0.719
----------------------------------
```

The reduction in the R^2 for the regression for the second year is caused by the limited trend component of the total labor (TL) series in the second year.

Relative effectiveness of rookies and fraction of experienced personnel for training. The relative effectiveness of rookies (ε) and the fraction of time they

102

require of experienced personnel for training (η) were established from the interviews. The range for initial effectiveness was estimated by an assistant manager – who has supervised at least 10 newcomers to the center – to be "about one-eighth [12.5%] for a person without lending experience to 75% for a person that just had to learn the new system."[17] The values reported by most lending officers were between 25% and 40%, with the exception of one that felt fully effective since day one. The adopted value for the relative effectiveness of rookies for the model is 35%.

The interviewees were much more consistent in recalling the time required of experienced personnel to supervise a rookie. On average, they thought that they required one hour of supervision/guidance per day during the first month. To make that estimate consistent with the extended training period in the model, the fraction of time required from experienced personnel for training (η) was set to 5%[18]. However, since informal guidance and question answering are omitted from this estimate, the 5% figure is likely to be an underestimate. Figure 4.5 shows the average gains in productivity from a rookie as estimated from the parameters obtained in the interviews.

To test the sensitivity of the effective labor fraction (e) to the parameters obtained from interviews – time for experience (τ_p), relative effectiveness of rookies (γ), and fraction of time required from experienced personnel for training (η) – a batch of Monte-Carlo simulations was performed. The three parameters were assumed to have a uniform distribution around their mean, and the range was set to cover at least 75% of the estimates out of the interviews. The simulations were run with all the sector equations active and driven by the historical authorized labor (AL). Figure 4.6 shows the results of those simulations with 50% and 95% confidence bounds.

Table 4.1 shows the effective labor fraction (e) at week 52 – after the stabilization of the hiring rate and before the convergence of all simulations – for the base case and extreme estimates of time for experience (τ_p) and relative effectiveness of rookies (γ). Changes in the fraction of time required from experienced personnel for training (η) within the specified range generated variations smaller than 1%.

Employee effectiveness

Employee effectiveness ————————————

Figure 4.5   Estimated learning curve
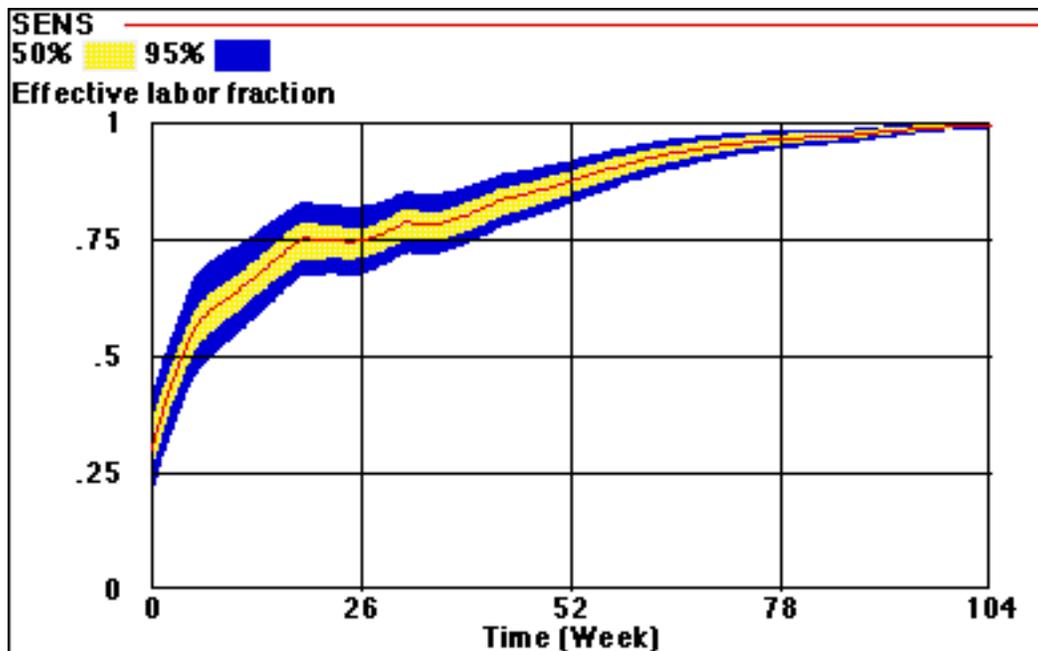
Figure 4.6   Sensitivity Analysis. Effective labor fraction

```
Number of runs = 100
t for experience          = RANDOM_UNIFORM [8,16]
Rookies' effectiveness    = RANDOM_UNIFORM [0.25,0.45]
Frac t for training       = RANDOM_UNIFORM [0.02,0.08]
```

Rookies' effectiveness	time for experience		
	8	12	16
0.25	0.910	0.856	0.803
0.35	0.921	0.874	0.828
0.45	0.931	0.892	0.852

Table 4.1   Sensitivity analysis. Effective labor fraction at week 52

§4.4.1.2. Capital Sector

A) Sector structure

The capital sector captures the firm's acquisition and disposal of capital resources through a stock-management system (Sterman, 1989b). Two co-flows track the technological content and required labor intensity of the firm's capital resources (Sterman, 1981). The main inputs to the sector are desired capital (K^*), and the exogenous capital technological content (ϕ) and capital labor intensity (θ) (see eqs. 17–28 in §3.5). Figures 4.7 and 4.8 show the stock and flow structure for this sector.
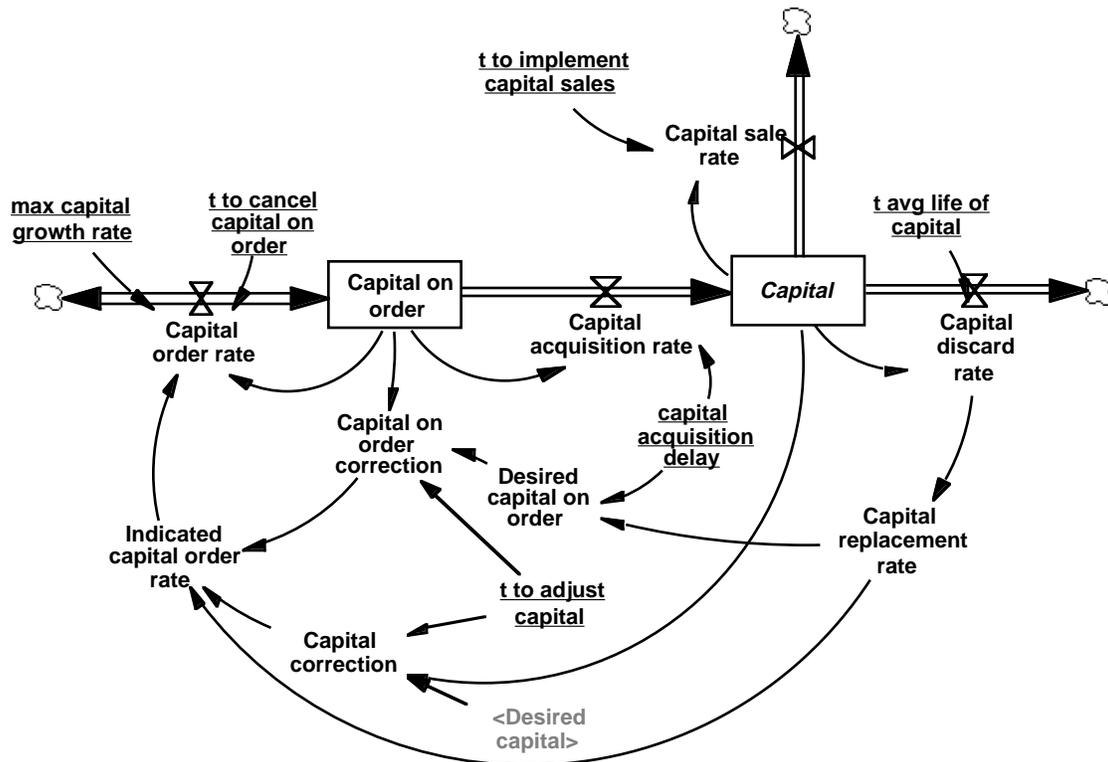
Figure 4.7   Structure of capital sector

Figure 4.8   Structure of capital sector (co-flows)

The following parameters need to be estimated for this sector:

- Time to adjust capital (τ_k)
- Capital acquisition delay (λ_k)
- Time to cancel capital in order (τ_s)
- Average capital life (τ_d)
- Time to implement capital sales (τ_{kd})
- Maximum capital growth rate (φ)
- Capital technological content (ϕ)
- Capital labor intensity (θ)

B) Parameter estimation

No information about the acquisition process for the capital equipment was accessible other than "computers have been available at the time new people are brought in."[19] Since technology has remained constant since the creation of the service center, and sufficient capital resources have been available for the employees, this sector is not a significant source of the dynamics observed in the service center. Parameters for this

---

[19] DR 6/27/95

sector were selected so as not to interfere with the main dynamics of the labor sector based on the assumption of Leontief technology.

To simplify the calibration process and the interpretation of results, the capital labor intensity (θ) was normalized at one. That is, one capital unit was fixed to represent the capital required for an employee to do his/her job – computer, telephone, desk, etc. The capital technological content (ϕ) was set to represent the normal productivity per work week of an experienced employee ($\phi = 35$ hours/week/capital). The stocks associated with the co-flows to track these characteristics were initialized assuming that existing capital was homogeneous with the same characteristics (a reasonable assumption given that the service center was established at the beginning of the simulation). The single-cohort structure for the co-flows was considered appropriate for this case.

Because of the required one-to-one relationship between employees and capital resources, the desired labor series was used to define desired capital. Parameters and initial conditions, however, were set to give an advantage to the capital acquisition process over the labor hiring process. Specifically, the capital acquisition lag (λ_k) was assumed to be shorter than the hiring delay (λ_l), and initial capital (K_0) greater than the initial total labor (R_0+E_0) – not unreasonable assumptions since the new systems were developed in advance and additional equipment could be added to the existing infrastructure without significant delays. Finally, the average capital life (τ_d) was assumed to be 10 years. Table 4.2 shows the values of the capital sector parameters and their equivalent in the labor sector.

Capital Sector	**Value**	**Labor Sector**	**Value**
Time to adjust capital (τ_k)	1 week	Time to adjust labor (τ_l)	1 week
Capital acquisition lag (λ_k)	6 weeks	Hiring delay (t to hire) (λ_l)	7.98 weeks
Time to cancel capital on orders (τ_s)	1 week	Time to cancel vacancies (τ_v)	1 week
Initial Capital (K_0)	40 cap. units	Initial Total Labor (R_0+E_0)	20 emps.
Average capital life (τ_d)	520 weeks	Time for turnover (τ_t)	400 weeks
Max. capital growth rate (φ)	6%/week	Max. labor growth rate (χ)	6%/week

Table 4.2   Parameter values for capital and labor sectors

Figure 4.9 shows the levels for capital and total labor resulting from a simulation with those parameters. The figure shows the availability of capital for employees at all times.

Since capital is assumed to be fully effective from its acquisition, this calibration does not interfere with the net service capacity available.
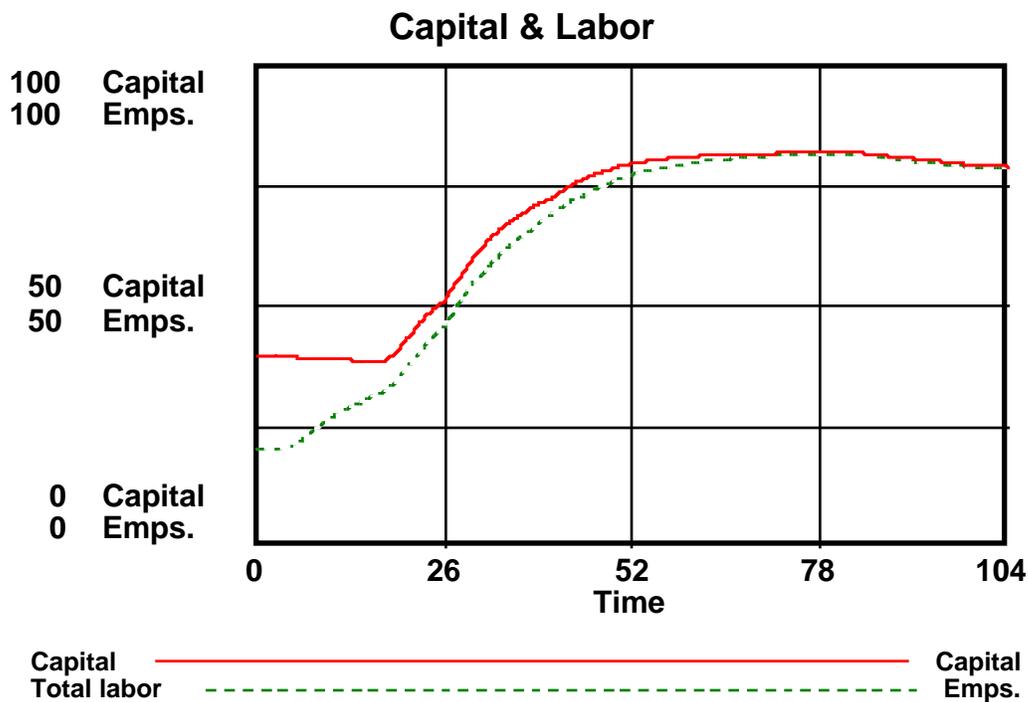
Capital & Labor

Figure 4.9   Capital and total labor (simulated data series)

Since the desired service capacity is hypothesized to respond to the desired throughput and time per order, the estimation of the parameters controlling the demand for the production factors will be presented after the estimation of the service delivery sector.

§4.4.2. Service Delivery

A) Sector structure

The service delivery sector accounts for customer orders as they flow through the service center. The desired order fulfillment rate is set to maintain a constant delivery delay. The sector also captures the employees' response to changes in work pressure and the effects of sustained work intensity – fatigue and burnout. The main inputs to this sector are customer orders (s_o), the nominal service capacity (sc_n), and the effective labor fraction (e) (see eqs. 32–46 in §3.5). Figure 4.10 shows the stock and flow structure for this sector.
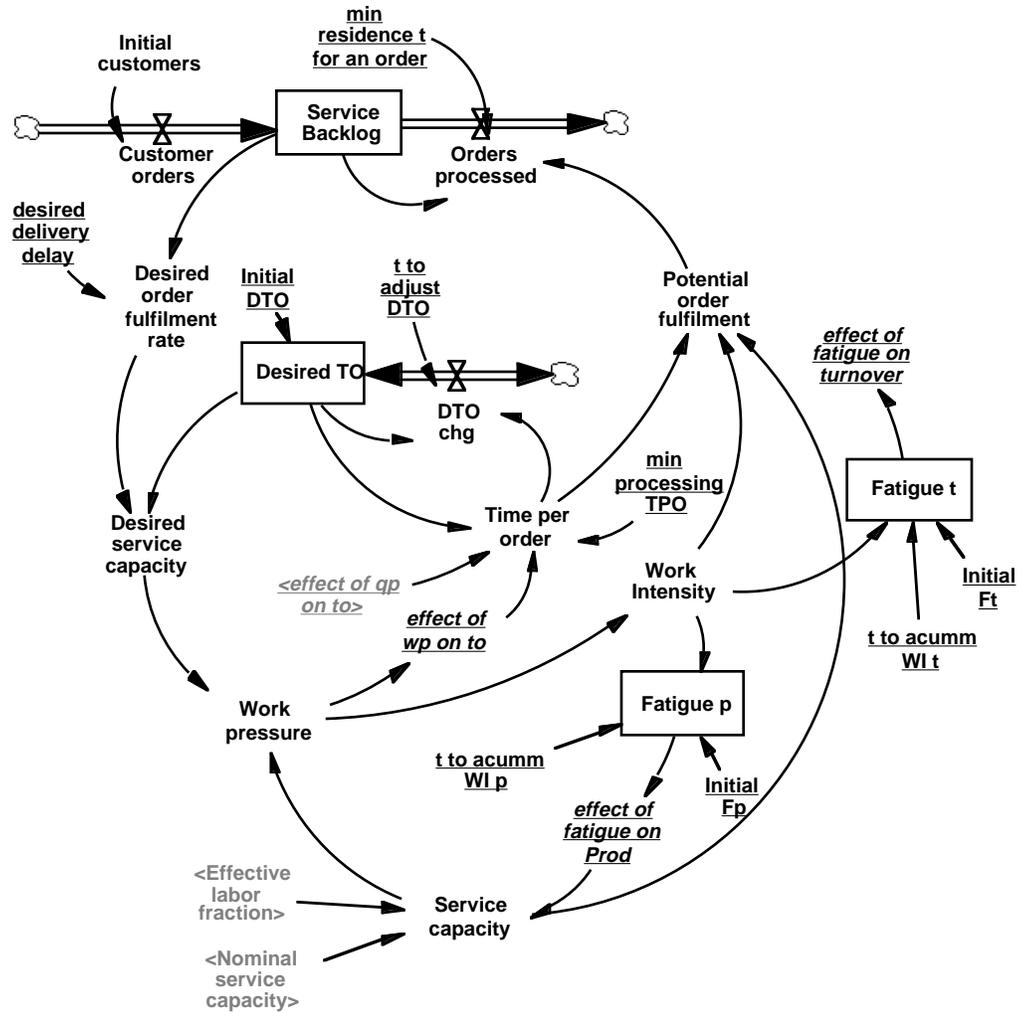
Figure 4.10   Structure of service delivery sector

In addition to the initial conditions for all the state variables, the following parameters
and non-linear functions need to be estimated for this sector:

- Minimum residence time for an order (τ_r)
- Minimum processing time per order (τ_p)
- Desired delivery delay (λ_s)
- Time to adjust desired time per order (τ_{to})
- Effect of work pressure on time per order ($t_w = f(p_w)$)
- Effect of work pressure on work intensity ($wi = f(p_w)$)
- Effect of fatigue on productivity ($efp = f(F_p)$)
- Time to fatigue for effect on productivity (τ_{fp})
- Effect of fatigue on turnover ($eft = f(F_t)$)
- Time to fatigue for effect on turnover (τ_{ft})

B) Data available

To estimate the response of lending officers to the changes in work pressure several operational data series were used:

Time Available (TA):   Total person-hours available to perform work in the Lending Center. The series was calculated from the weekly Branch Operating Reports for the June '94 to May '95 period [person-hrs/week].

```
TA= Paid full time - Holiday hrs - Sickness hrs -
    (14 * Days in week)
```

The subtraction of the equivalent of two manager work-days is consistent with the practice of having somebody cover in case a manager is absent because of vacation or sickness.

Time to Process (TP):   Total person-hours allocated to processing customer orders. The series was calculated from the weekly Branch Operating Reports for the June '94 to May '95 period [person-hrs/week].

```
TP= Time available + Paid overtime + Net relief hrs
    - Meeting hrs - Training hrs
```

Order Fulfillment (OF):   Total number of customer orders processed per week in equivalent standard hours. The series was directly available from the weekly Branch Operating Reports for the June '94 to May '95 period [orders/week].

```
OF= Variable earned hrs
```

Finally, the effective labor fraction (e) estimated in the labor sector was used to capture the effective productivity of labor in full-time equivalents of experienced personnel (ELF).

Since orders can only arrive to the LC if it is open, the main determinant of how many orders are processed per week (OF) is the number of days that the lending center is open, i.e., time available (TA). Figure 4.11 shows the two data series, where the large dips below the normal operating point represent weeks with one or two days off due to holidays. As it would be expected, there is significant correlation between the number of orders processed (OF) and the time available to process them (TA) ($r=0.87$) – the number of hours available in any particular day limit the amount of orders that could be processed. However, normalizing the two series for the number of days in the work week,

and removing outlier points[20], the correlation between the two series is significantly reduced (r=0.22), thus warranting an exploration of the structure of how employees respond to changes in work pressure.

| 3,000 | hrs/week |
| 3,000 | ords/week |
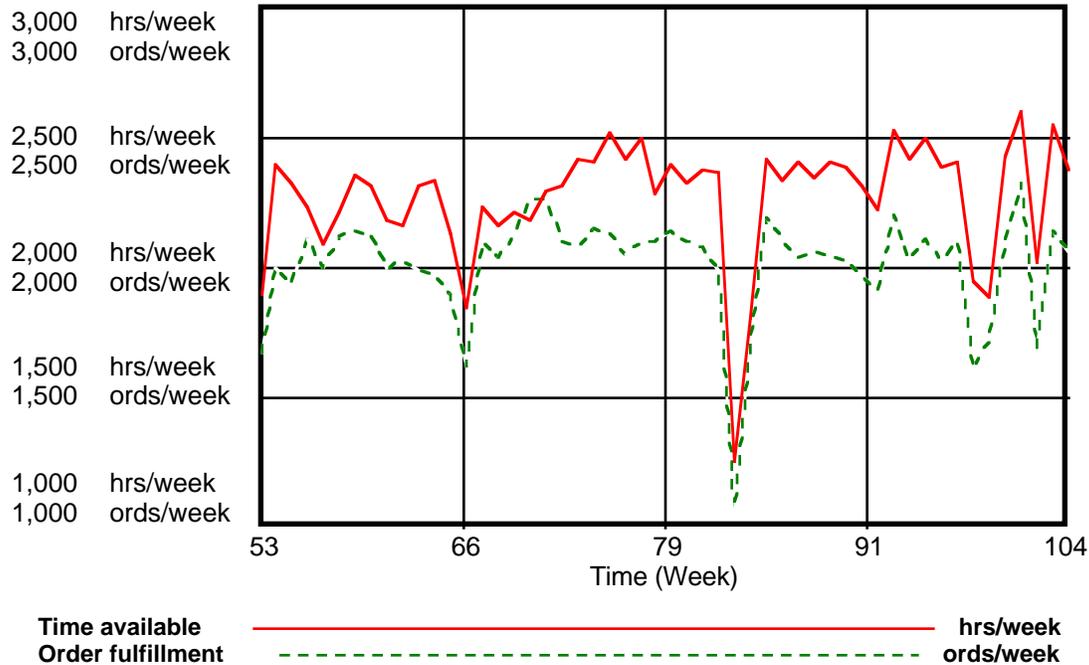
Figure 4.11   Time available and order fulfillment (historical data series)

C) Parameter estimation

`Minimum time to process an order and desired delivery delay.` From observation and interviews it was possible to determine that there is no significant backlog from one week to the other in the service center. As one of the lending officers stated: "I don't like pending so I try to answer post, whatever, all in the same day."[21] In general, phone requests, incoming mail, and exception reports must be processed the same day. The only activities that have an extended delivery delay are the investigations and responses triggered by the quarterly reports used to monitor the risk grade and credit limits of the portfolio; these activities, however, represent less than 4% of the overall work load for loan officers. Consistent with these observations, the minimum residence

---

[20] Two outliers from the normal operating point were identified: Christmas week, and a week where there was an identifiable error in the collection of throughput data.

[21] AL 6/29/95

time for an order (τ_r) and the desired delivery delay (λ_s) were fixed at half a day [0.1 weeks].

Because of the same-day turnaround and the minimum backlog held at the service center, the rate of incoming customer orders (s_o) is assumed to be equal to the rate of outgoing orders (OF). The backlog stock was initialized to the operating backlog in equilibrium ($s_o*\tau_r$).

The effective service capacity that is applied to processing orders is given by modifying the nominal service capacity, the production function, by the effects on labor productivity, i.e., experience and fatigue. Because of the availability of capital resources for all employees, the total number of person-hours available to perform work in the lending center – time available (TA) – will be used as a proxy for nominal service capacity (sc_n). The effects of experience in productivity have already been captured in the effective labor fraction (e). To determine work pressure, it is first necessary to estimate the effects of fatigue on labor productivity.

Effects of fatigue on labor productivity and employee turnover. From the original data sources, it was possible to estimate the weekly work intensity response of the lending center (WI=TP/TA). The dimensionless ratio can be conceptualized as the fraction of time available allocated to processing orders. However, because of overtime and relief hours, this number can be greater than one. By inspection of the work intensity time series (figure 4.12), it becomes evident that its range is within the 90 to 105 percent for a standard work week of 35 hours [33.25 to 36.75 hours].

Studies in the manufacturing and construction industry have shown that productivity decreases when the work week exceeds 45 hours for more than three weeks (Kossoris, 1947). For burnout to have an impact on employee turnover, it has to be sustained for longer periods of time (Golembiewski, Munzenrider and Carter, 1983; Pines and Kafry, 1978). Thus, the data available are not in the range that can be used to estimate the actual loss of productivity or increase in turnover because of sustained work intensity. For completeness, the effect of fatigue on productivity (*efp*) table function was estimated using information from the manufacturing industry; adjusting it for a normal work intensity of 35 hours per week (see figure 4.13). To be consistent with the data for the manufacturing sector, the time constant to smooth the work intensity (τ_{fp}) was set to three weeks.

Work Intensity

Figure 4.12   **Work intensity (historical data series)**
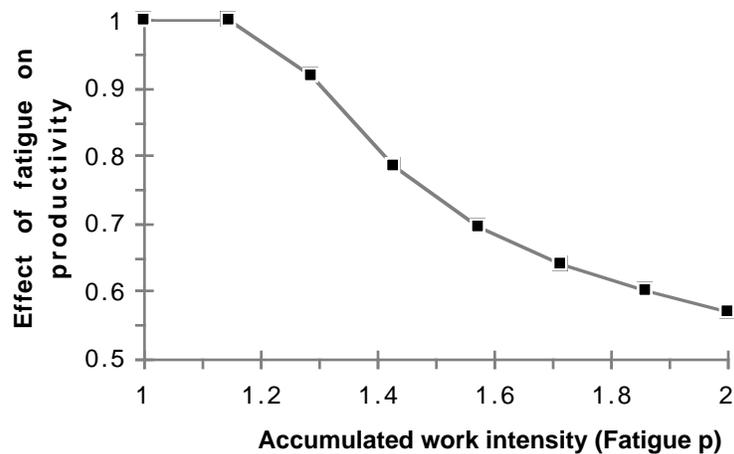
Figure 4.13   **Table function. Effect of fatigue on productivity**

The effect of fatigue on turnover (*eft*) table function was estimated using information from the human service organizations (e.g., police officers, social workers, nurses and teachers). Most studies done in this area estimate the effect of burnout on the intention to leave (Kirschenbaum and Weisberg, 1990; Parasuraman, 1982). Jackson, Schwab and Schuler (1986) have shown that, although burnout is strongly linked to intentions to

leave, these intentions might not be translated into action since the process is complex and includes intermediate linkages and environmental variables. The formulation proposed in equation eight takes this into consideration by assuming a normal turnover time for the industry, and modifying the responsiveness to it with the effects of fatigue and perceived quality.

Figure 4.14 shows the calculated effect of burnout on intention to leave as derived by Weisberg's (1994) physical exhaustion. The time constant to smooth the work intensity for the effect on turnover was estimated at 52 weeks based on the longitudinal study done by Jackson, Schwab and Schuler (1986). Although the response estimated by Weisberg is linear, the exponential accumulation of work intensity introduces the non-linear response suggested by the phases of progressive burnout suggested by Golembiewski et al (1983).
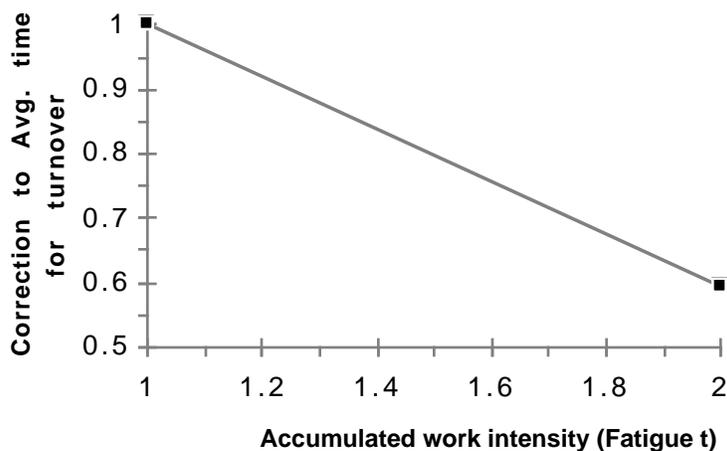
Figure 4.14   Table function. Effect of fatigue on turnover

Having determined the effect of fatigue on productivity (*efp*) the following transformations were done to the original data available for the estimation of the work pressure.

Effective Time Available (ETA):

> Correction of time available by the effective labor fraction (ELF) and the effect of fatigue on productivity (*efp*). This corresponds to correcting for the effects of experience and fatigue on labor productivity [person-hrs/week]. The series was used as service capacity under normal work intensity.
>
> `ETA= TA*(ELF)*(efp)`

Time per order (TPO):   Effective time required to process a standard order. The time to process is corrected for the effects of experience (ELF) and fatigue (*eft*) and divided by the number of orders processed per week [person-hrs/order].

```
TPO= TP*(ELF)*(efp)/OF
```

Time per order. The desired time per order (T^*) and work pressure (p_w) are tightly coupled through two feedback loops (see figure 4.10). The first loop generates the 'anchoring and adjustment' process around time per order (T) and desired time per order (T^*). A second loop is formed by desired time per order (T^*), the desired service capacity (sc^*), work pressure (p_w), and time per order (T). Because of these interdependencies, the set of equations had to be estimated together.

The function for the effect of work pressure on time per order (t_w) is hypothesized to have a value of one when work pressure is neutral ($p_w=0$) and be monotonically decreasing (see eq. 43 in §3.5). For estimation purposes an exponential function was tested for this formulation:

$$t_w = f(p_w) = Exp(\alpha p_w)$$

Under this formulation the coefficient of the effect of work pressure on time per order (α) is expected to be negative. Figure 4.15 shows the shape of the effect of work pressure on time per order under a range of values for alpha.
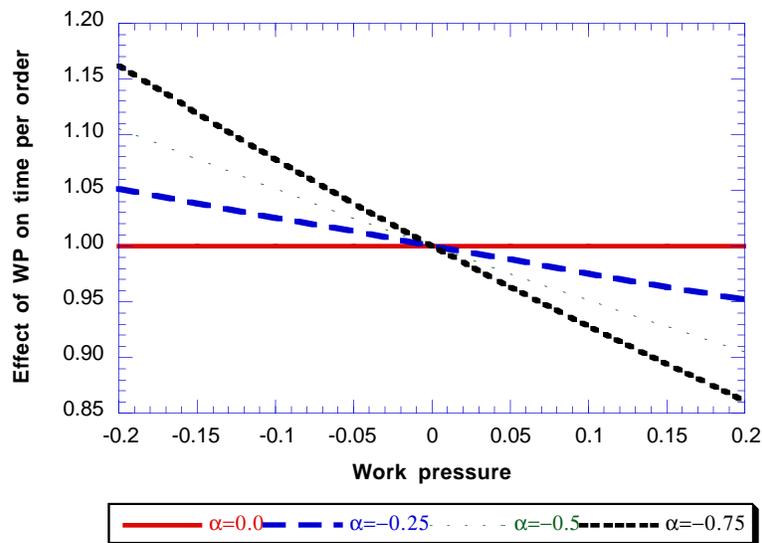
Figure 4.15   Sensitivity to alpha. Effect of work pressure on time per order

Non-linear least squares estimation was used to determine the coefficient (α) for the effect of work pressure on time per order (t_w), time to adjust desired time per order (τ_{to}), and an initial value for the desired time per order (T_0^*).

$$\underset{T_0^*,\alpha,\tau_{to}}{Min} \sum_{t=53}^{104}\left(T(\tilde{t}) - TPO(\tilde{t})\right)^2 \qquad \text{for}\left\{\tilde{t} \mid \tilde{t} \in t \text{ and } TPO(t) = value\right\}$$

Subject to

$$T(t) = t_w(t) \cdot T^*(t)$$

$$T^*(t) = T^*(t - dt) + dt\left(\left(T(t) - T^*(t - dt)\right)/\tau_{to}\right)$$

$$sc^*(t) = OF(t) \cdot T^*(t)$$

$$p_w(t) = \left(sc^*(t) - ETA(t)\right)/ETA(t)$$

$$t_w(t) = Exp\left(\alpha p_w(t)\right)$$

The results of this estimation gave reasonable values for alpha and initial desired time per order but a very large value for time to adjust desired time per order (τ_{to}), i.e., no changes in the desired time per order. The residuals of this estimation, however, showed a significant downward trend, pointing to an erosion of the desired time per order.

Alternative formulations and initial conditions were tested, but not until the constraint for symmetrical adjustment of the desired time per order (T^*) was removed did the trend in the residuals disappear. The main characteristic of non-symmetrical adjustment processes is that they permit different speeds of adjustment for the underlying level, depending on the direction of the adjustment. Non-symmetrical adjustment processes have been used in the organizational and psychological literature to represent the biased formation of expectations and goals (Lant, 1992). The non-symmetrical adjustment is normally formulated by allowing different time constants to govern the adjustment process depending on whether the aspiration level is above or below the actual performance. For estimation purposes, an additional equation for time to adjust desired time per order (τ_{to}) was introduced:

$$\tau_{to} = \begin{cases} \tau_{up} & \text{If}(T(t) > T^*(t)) \\ \tau_{down} & \text{otherwise;} \end{cases}$$

The estimation process with this formula yielded a time constant to adjust desired time per order upward (τ_{up}) greater than 43 years, that is, effectively infinite, while allowing the desired time per order to drift down with a time constant of less than 20 weeks.

116

The effects of quality pressure on time per order were not introduced in the original estimation of desired time per order because of difficulties in obtaining a data series that could be used for this purpose. The difficulty in obtaining an operational metric is in itself data to support the proposed hypothesis for erosion of service quality, i.e., lack of operational metrics of service quality. However, when such a long time constant was identified for the upward adjusting process it became evident that no such upward pressures were in place in the operations of the lending center, and efforts to identify such a data source were abandoned.

To reduce the number of estimated parameters I decided to make the adjustment of desired time per order exclusively downward by modifying the rate of change of desired time per order to:

$$(d/dt)T^* = \min\left(0, \left(T - T^*\right)/\tau_{to}\right)$$

Estimating the updated structure generated the following results with a 95% confidence interval:

```
13.2555 <= T TO ADJUST DTO = 18.7321 <= 28.7598
-0.6915 <= ALPHA =          -0.6393 <= -0.5863
 1.0591 <= INITIAL DTT =     1.0758 <=  1.0933
```

All the estimates have tight confidence intervals. As expected, alpha has a negative sign thus reducing the time per order when work pressure is high. From the work intensity data (see figure 4.12), we can tell that the operations of the center are close to the normal operating point, thus confidence in the estimated response should drop for values of work pressure beyond this range.

The estimate for the initial desired time per order ($T_0{}^*$) and the values that desired time per order (T^*) takes during the simulation (see figure 4.16), seem to be low for the bank's productivity standard of one hour of preparation and breaks for every six hours dedicated to processing orders – an implied T^* of 1.166. The estimate, however, is consistent with data obtained from interviews stating that service personnel tend to work unreported overtime.

> ... 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].[22]

---

[22] AL 6/29/95

> ... and they don't always claim that over time either ... I suppose that they're worried that someone would say "you are not working very clever" (sic) or something. And then the assistant managers, they don't get paid over time ... I never go out to lunch; I'm giving the bank five hours a week of overtime.[23]

The analysis of residuals showed that they were normally distributed, with mean zero and no significant lags in the autocorrelation function (ACF). Only one point of the partial autocorrelation function (PACF) (Wei, 1990) was significant (lag 15), but one out of 20 points is to be expected with 95% confidence intervals. The fit between the simulated series and the series estimated from the data does not show any systematic error:

Time per Order

TPO simulated	————————	**hours/task**
TPO historical	- - - - - - - - - -	**hours/task**
Desired TPO	— — — — — —	**hours/task**

Figure 4.16   Time per order (historical and simulated data series)

---

[23] CK 6/28/95

```
Summary Statistics for Historical Fit
Time per Order
---------------------------------
n =   50.0

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

Unfortunately, it is impossible to determine from the data available whether the calculated time per order is indeed allocated to processing customer orders or whether employees just allocate the desired time per order and spend the rest of the time in unproductive tasks. To explore this issue I decided to test if the response to work pressure on time per order (the α coefficient) changed from times of high work pressure to times of excess capacity. It is expected that if there is excess capacity the response to work pressure on time per order should be less aggressive than in times when work pressure is high.

From the work intensity data (see figure 4.12) it is clear that the center had two distinctive operating levels during the year when data are available. From week 52 to week 73 the work intensity in the lending center was above the normal operating point. After week 74 there is a consistent over capacity in the lending center. The threshold for the two operating points is consistent with the aggressive hiring rate seen before week 74, and the sudden drop of personnel hiring seen thereafter (see figure 4.3).

Non-linear least squares estimation was used to determine the coefficient (α) for the effect of work pressure on time per order (t_w), time to adjust desired time per order (τ_{to}), and an initial value for the desired time per order (T_0^*) with the data from each of the operating points. The results, with 95% confidence intervals, were as follows:

```
Under work pressure -> From week 52 to week 73:
   6.4669 <= TAO =        11.2245 <= 23.4583
  -0.7442 <= ALPHA =       -0.6607 <= -0.5827
   1.0585 <= INITIAL DTT = 1.0907 <=  1.1274

With excess capacity -> From week 74 to week 104:
   0.6314 <= TAO =         1.7473 <= 47909.5
  -0.6108 <= ALPHA =       -0.5402 <= -0.4676
   1.0483 <= INITIAL DTT = 1.0631 <=  1.0811
```

The estimates for the first section of the data are statistically equal as the ones estimated for the whole data series. For the period with excess capacity, the estimator for the time to adjust desired time per order (τ_{to}) is unreliable (see the confidence interval) because there is no significant variation in the dependent variable to estimate it. The rest of the estimates have reasonable tight confidence intervals.

Although the coefficient that captures the response to work pressure is slightly less aggressive with excess capacity, there is no evidence to reject the null hypothesis that the alpha coefficient does not change when there is excess capacity – the partial α coefficients are not statistically different from the α coefficient estimated for the full series. Even though it is impossible to determine the actual allocation of time for processing each order, it seems that employees are consistent in their response to work pressure. Further support to the idea of full utilization of excess capacity in processing orders was found when estimating the effects of time per order on sales (see § 4.3.3).

Having estimated the work pressure (p_w) that employees of the lending center were subject to it is now possible to determine the responsiveness of work intensity to work pressure.

Work Intensity. The response of work intensity (*wi*) to work pressure is hypothesized to have a value of one when work pressure is neutral, and to be monotonically increasing (see eq. 36 in §3.5) in work pressure. For estimation purposes a function similar to the response of time per order to work pressure (t_w) was used.

$$wi = f(p_w) = Exp(\beta p_w)$$

In this case, however, the coefficient for the response (β) is expected to be positive. Non-linear least squares estimation was used to determine the coefficient. Work pressure was derived using the estimates obtained for the desired time per order.

$$\underset{\beta}{Min} \sum_{t=53}^{104} \left(wi(\tilde{t}) - WI(\tilde{t}) \right)^2 \qquad for\{\tilde{t} \mid \tilde{t} \in t \text{ and } WI(t) = value\}$$

Subject to

$$wi(t) = Exp\left(\beta p_w(t)\right)$$

$$p_w(t) = \left(sc^*(t) - ETA(t)\right)/ETA(t)$$

$$sc^*(t) = OF(t) \cdot T^*(t)$$

$$T^*(t) = T^*(t - dt) + \min\left(0, dt\left(\left(T(t) - T^*(t - dt)\right)/\tau_{to}\right)\right)$$

$$T(t) = t_w(t) \cdot T^*(t)$$

$$t_w(t) = Exp\left(\alpha p_w(t)\right)$$

$$\tau_{to} = 14.64 \qquad \alpha = -0.648 \qquad T_0^* = 1.086$$

The estimator for beta has the correct sign and a tight confidence interval, i.e., it is significantly different from zero.

```
0.3201 <= BETA =   0.3705 <=   0.4211
```

Residuals of the estimated equation are normal with mean zero and with no significant patterns in the ACF and PACF plots. Although the fit to the historical series is not as good as the one for time per order, it does not show a systematic error.

```
Summary Statistics for Historical Fit
Work Intensity
-----------------------------------
n =   50.0

R^2                          0.667
Mean Abs. Percent Error      0.016
Mean Square Error            0.000
Root Mean Square Error       0.019
Bias                         0.023
Variation                    0.072
Covariation                  0.904
-----------------------------------
```

Figure 4.17   Work intensity (historical and simulated data series)

Comparing the intensity of the two possible responses that employees have to a change in work intensity (see figure 4.18), it is possible to see that there is a more aggressive response to reduce the time allocated to process each customer order than the response to increase the work intensity ($|\alpha| > |\beta|$). For example, if work pressure is 0.1, then time per order would be reduced by 6.2%, while work intensity would only increase 3.7%. The prioritization of responses is consistent with the fact that the center has been operating with excess capacity since week 74. Excess capacity has allowed employees to allocate more time to process each customer order than the desired level (see figure 4.16), thus mitigating their objections to reduce that time when work pressure goes up. However, as shown in the estimation of the effect of work pressure on time per order, the response priority does not seem to change when the employees are under work pressure, hence the potential for erosion of service quality.

Responses to Work Pressure
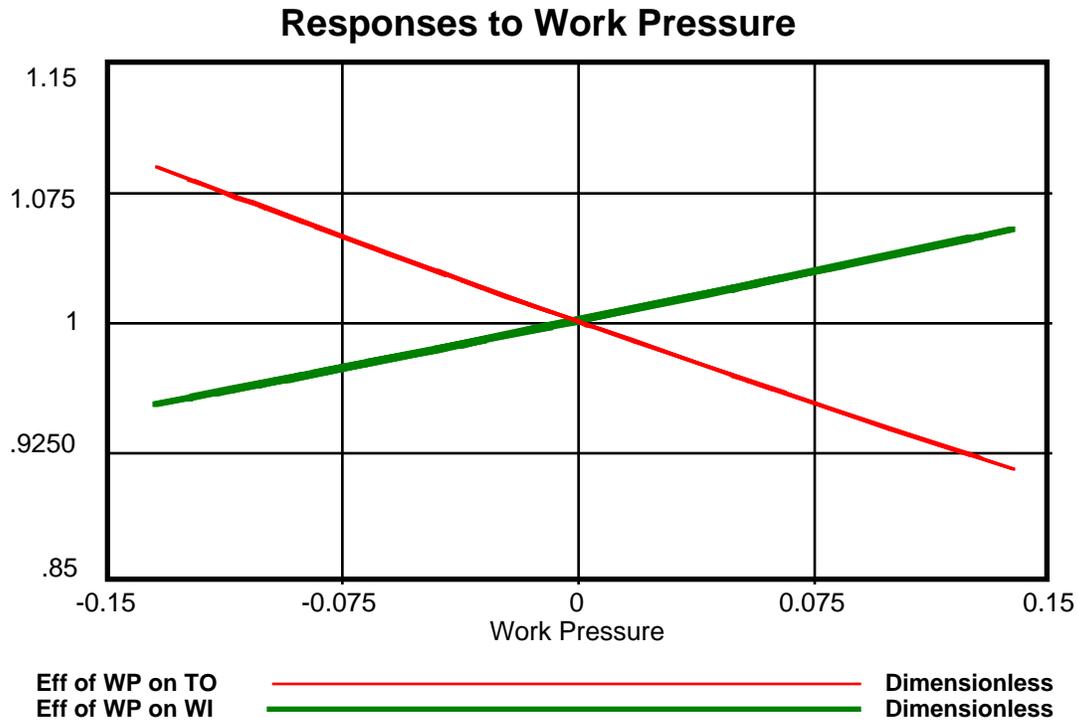
Figure 4.18   Estimated employees' responses to work pressure

§4.4.3. Factor Demand

A) Sector structure

This subsector depicts management policies for establishing the desired production factors as a response of the incoming customer orders and the desired time per order (see eqs. 29–31 in §3.5). In the generic model presented in the previous chapter, the optimal estimation of requirements of production factors was presented as an *a fortiori* assumption. In this section, however, I will develop a more realistic formulation to capture the observed dynamics in the lending center.

B) Data available

Four data series were used to determine the policies that regulate the authorization of personnel:

Authorized Labor (AL):   Number of employees authorized to work in the Lending Center. The time series for the first year of the center's operations (June '93-June '94) was obtained from a study done by the center's

manager. The Branch Operating Reports provided monthly data for the June '94 to May '95 period (employees).

Total Labor (TL):   Total number of employees working in the Lending Center. The series was directly available from the monthly payroll for the first year of the center's operations (June '93-June '94) and from the weekly Branch Operating Reports for the June '94 to May '95 period (employees).

Effective Service Capacity (ESC): The effective service capacity estimated in the service delivery sector was used to capture the effects of experience, fatigue and absenteeism on productivity. Data were made available on a weekly basis (effective hours/week).

Desired Service Capacity (DSC): The desired service capacity estimated in the service delivery sector was used to capture the desired processing capacity required by the lending center (effective hours/week).

Since there was no information available on customer demand during the buildup phase of the lending center, it was not possible to estimate the desired service capacity for that period. However, using data from the center's second year of operations – by which time most of the branches had been absorbed – yielded a more realistic estimation of the operating policies under normal circumstances.

C) Parameter estimation

Because of the assumption of Leontief technology, i.e., the inelastic substitution of production factors, and the lack of data on the capital sector, no attempt was made in this calibration to capture the effects of price on the demand of the production factors.

For consistency with the base model presented in Chapter 3, it was presupposed that management perceives the effectiveness of the labor force after a fixed time delay (see eq. 29 in §3.5). However, two changes were introduced to make the policies more realistic. First, absenteeism was introduced as an additional factor for loss of productivity. Labor effectiveness was calculated as the ratio of effective service capacity to nominal service capacity. Since the effective service capacity was obtained from the number of employees that showed up to work, absenteeism was automatically integrated. Second, instead of assuming perfect information and reactiveness in the acquisition policy, it is

postulated that management smoothes the response from weekly calculations of indicated desired labor.

Assuming a desired level of production, the cost-minimizing factor demand for Leontief technology is computed as the division of the desired output by the productivity of a unit of a production factor. Both the capital's average technological index (*a*) and the average labor intensity (*i*) were assumed constant for the period for which data was available. Since desired service capacity was estimated in effective hours per week, the average technological index was set at 35 hours/week/capital, the center's normal work-week, while the average labor intensity was kept at its assumed value of one.

Non-linear least squares estimation was used to derive the initial value of perceived employees effectiveness (E_0), the time constants for the perception of employees effectiveness (τ_{pe}), and the time to adjust desired labor (τ_{l*}).

$$\underset{E_0,\tau_{pe},\tau_{l^*}}{Min} \sum_{t=53}^{104}\left(L^*(\tilde{t})-AL(\tilde{t})\right)^2 \qquad \text{for}\left\{\tilde{t} \mid \tilde{t} \in t \text{ and } WI(t)=value\right\}$$

Subject to

$$L^*(t)=L^*(t-dt)+dt\left(\left(l(t)-L^*(t-dt)\right)\Big/\tau_{l^*}\right)$$

$$l(t)=\frac{DSC(t)\cdot i}{a\cdot E(t)}$$

$$E(t)=E(t-dt)+dt\left(\left(le(t)-E(t-dt)\right)\Big/\tau_{pe}\right)$$

$$le(t)=\frac{ESC(t)}{TL(t)\cdot(a/i)}$$

$$a=35 \qquad i=1$$

The resulting estimators have the correct sign and are significantly different from zero.

```
 4.6975 <= T TO PERCEIVE LE = 6.7026 <=  9.3945
 0.7755 <= INITIAL PELF =      0.7790 <=  0.7823
16.2665 <= T TO ADJUST DP =   18.7602 <= 21.6765
```

Although the data do not allow the two time constants to be effectively separated – an intermediate data series would be required – the value estimated for the perceived labor effectiveness at time 52 ($E_0 = 0.77$) is consistent with the exponential smoothing of the data available for that series with the time to perceive labor effectiveness (6.7 weeks) as the time constant.

The theory proposed in Chapter 3 suggests that management does not have a good appreciation of the intangible metrics of service capacity, e.g., effects of fatigue and experience on productivity, hence a high initial perceived labor effectiveness and a long time constant to correct it were expected. The estimated time to perceive labor effectiveness (6.7) does not match the theory's prediction. Two elements were confounded in the perception of labor effectiveness in the LC that seem to increase management's awareness of labor's productivity. First, there were no reliable estimates of labor productivity for the Nelson House lending center. Although some lending centers had been operating before Nelson House started operations, the service-mix that those centers were handling was different from the mix present in the branches of London West End. Productivity estimates had to be developed during the first year of operations of the lending center, when most employees were yet to gain experience and many of the standard operating procedures were being debugged. The second factor affecting the perception of labor productivity is absenteeism. The rate of absenteeism is confounded with labor productivity in the estimation process, thus increasing the tangibility of the losses in productivity.

One of the managers I interviewed in the Regional Office thought that the adjustment of this perception might be biased. He felt that although management does immediately recognize gains in productivity because of better systems, debugging of processes, and overall experience with the systems, it never considers the employee's learning curve in the estimation of labor requirements. The data, however, do not allow separation of these effects[24].

Despite the apparent awareness of labor productivity, the overall responsiveness of the hiring policy seems to be rather slow – on top of the 18.7 weeks to adjust desired labor (τ_{l*}), there are nine weeks to adjust labor (τ_l), and additional nine weeks for the hiring process (λl) to take place. A possible explanation for the long estimate of the time to adjust desired labor (τ_{l*}) is that the estimate was based on a period when the lending center had excess capacity. The bank's tradition is not to lay off personnel and, according to an officer from corporate human resources, "the difficulties that the bank is having to

---

[24] An alternative solution to this difficulty would have been to estimate the perception of productivity as an asymmetric adjustment process. The simulation results, however, did not show any significant reductions in labor productivity – limited effects of fatigue and experience mix – to make the estimation reliable. §4.5.3 explores the implications of an asymmetric adjustment process.

manage careers under 'new deals' [incentives other that life employment]"[25] might have overestimated the time constant.

Notwithstanding these difficulties, the proposed structure seems to be a valid representation of the policies to determine the desired labor force that were in place during the second year of operations of the lending center. The comparison of the simulated series to the historical data does not show any systematic error.

```
         Summary Statistics for Historical Fit
         Desired Labor
         ----------------------------------
         n =   54.0

         R^2                         0.770
         Mean Abs. Percent Error     0.008
         Mean Square Error           0.800
         Root Mean Square Error      0.895
         Bias                        0.000
         Variation                   0.051
         Covariation                 0.949
         ----------------------------------
```

§4.4.4. Service Quality

A) Sector structure

This sector models the perceptions and expectations of quality for the three main agents involved in the service delivery process – employees, managers, and customers – as recognized by the service center (see eqs. 47–55 in §3.5). Quality perceptions are modeled as exponential adjustment processes of the actual quality delivered; each agent is assumed to have a different time constant for the adjustment process depending on its proximity to the service delivery process. The quality expectations for each agent are modeled as a level of aspiration, with exogenous anchoring to represent the intrinsic biases of each agent. A set of coefficients is used in each of these calculations to determine the relative weight of the various inputs to the formation of expectations. The main input to the sector is the time allocated per order (T), and the main outputs are the effects of quality pressure on time per order (t_q) and the effect of perceived quality on turnover (eqt). Figure 4.19 shows the stock and flow structure for this sector.

---

[25] PC 10/24/95
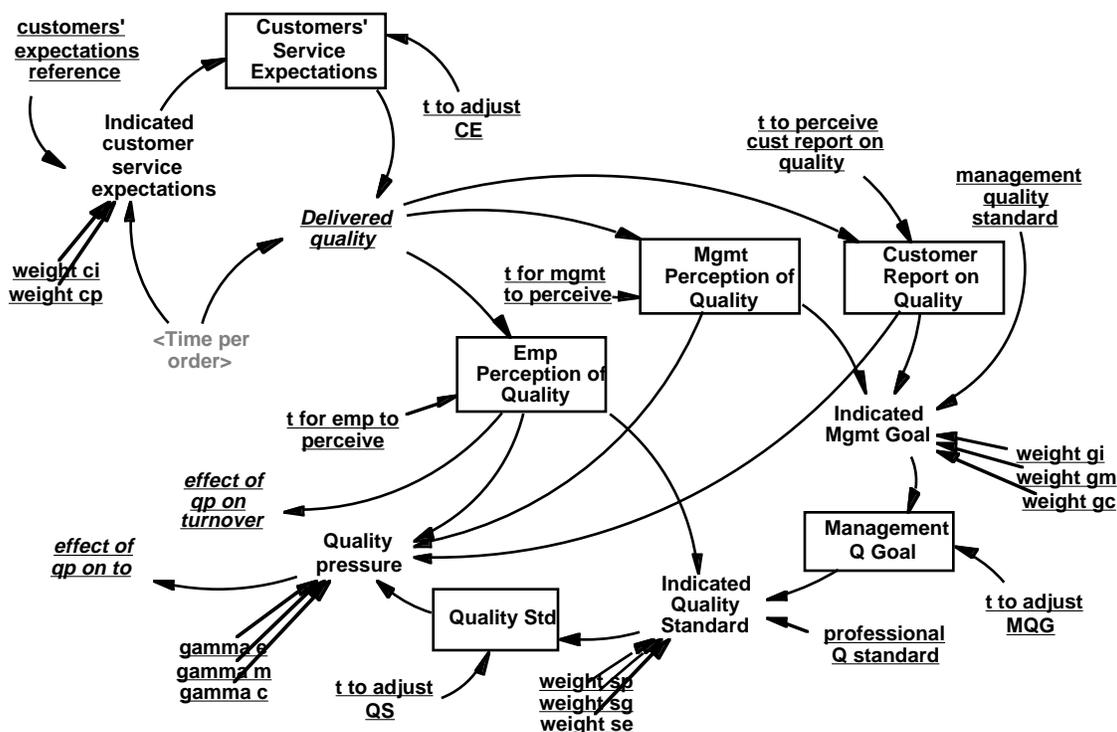
Figure 4.19   Structure of service quality sector

In addition to the initial conditions for all the state variables, the following parameters
and non-linear functions need to be estimated for this sector:

- Employees' bias in the perception of service quality (β_e)
- Management's bias in the perception of service quality (β_m)
- Customer's bias in the reporting of service quality (β_c)
- Employees' time to perceive quality (τ_{qe})
- Management's time to perceive quality (τ_{qm})
- Time to recognize customers' report on quality (τ_{qc})
- Weights for formation of employees' quality standard ($\alpha_{sp}, \alpha_{sg}, \alpha_{se}$)
- Weights for formation of management's quality goal ($\alpha_{gi}, \alpha_{gm}, \alpha_{gc}$)
- Weights for formation of customers' expectations (α_{ci}, α_{cp})
- Weights for formation of quality pressure ($\gamma_e, \gamma_m, \gamma_c$)

B) Data available

Three mechanisms are in place in the Lending center to monitor the quality of the service
delivery process. First, NatWest Retail Banking Services as a whole has an instrument to
monitor quarterly customer satisfaction. The Customer Satisfaction Index (CSI) is a 34
item questionnaire asking customers to evaluate the service they have received as

personal customers – as opposed to dealings involving business accounts – from the bank in the last six months. The survey is sent out quarterly, and the sampling methodology ensures that all customers in good standing, i.e., with no bad debt, receive the questionnaire once every three years. The number of quarterly responses between 1991 and 1994 has been between 80,000 and 100,000; a response rate between 20% and 25%.

More specific to the activities of the lending center are the Quality of Book statistics that are compiled monthly by the regional office. The main indicator from these statistics is the risk index of the lending book. The bank assigns a risk grade to each account based on the credit history of the of the loan holder – lower grade reflects lower risk. The Risk Index is calculated for the center's lending book – the total outstanding loans held by branches supported by the LC – by weighting the each account's risk grade by the loan amount.

Finally, the management of the Nelson House lending center has designed an instrument that is sent twice a year to the managers of the customer service branches being served by the lending center – the Lending Center Quality Service Monitor. The instrument is a questionnaire with eight questions asking to grade in a 1 to 10 scale the various services that the lending center provides for the customer service branch. The Lending Center Quality Service Monitor was introduced in February 1994, and, at the time of this study, three measurements were available.

Other than the Lending Center Quality Service Monitor, there are no specific operational metrics on service quality. All the measurements of daily operations are, as predicted by the theory, focused on throughput of customer orders.

C) Parameter estimation

Formation of quality standard and quality pressure. Since the quality sector captures how the agents use the quality information – if available – to modify the service delivery process, the discussion will be centered in the use of the quality reports described above.

Although the manager and assistant manager of the lending center were aware that the CSI data were being compiled, it had never been provided to them during their tenure in the lending center. The first access they had to the CSI data was when I showed it to them in an interview. In any case, the data collected by the CSI are of little use in the operations of the lending center. The questionnaire was designed with the traditional

customer service branch in mind, and most of the items in the questionnaire are about the operations and services in the branch holding the account. It is not possible to construct from the CSI data a profile of the backroom operations supporting the customer service branches (e.g., service center, lending center, etc.).

The Risk Index of the lending book is, by far, the most prevalent measure of performance for management in the lending center. Not only does the topic of the success of the LC on improving this metric come up within minutes of any interview, but a chart comparing the lending center's Risk Index to the regional average is prominently displayed in the manager's office. The metric reflects the impact of standardized decision process for the lending center, but it fails to capture the customers' satisfaction with the process.

Although the LC Quality Service Monitor was designed and implemented by the LC's management, the impact of the data collected through the instrument has been limited. Most of the indicators relating to responsiveness and usefulness of the service provided have shown an increase over time that matches the increase of time allocated per order (see figure 4.16). The problem with the instrument, however, is it that is not precise or frequent enough to be used as a real feedback instrument to adjust the service delivery process. When probed about the actions implemented in response to the indicators, the manager of the LC admitted that the information collected was neither reliable nor useful. None of the lending officers were aware of the instrument nor their performance as evaluated by branch managers.

Summarizing, management has limited instruments to assess customer satisfaction operationally, and, even in the areas where some gaps have been identified, no corrective action has been enacted.

Because of the lack of operational metrics of service quality as perceived by the customer – either the customer service branch or the final customer – it is safe to assume that the formation of the quality standard and quality pressure are exclusively driven by employees ($\alpha_{se} = \gamma_e = 1$). The results of a study done by the Region Human resource office in January 1995 provide further evidence for this assumption. In response to the open-ended question "What is your main area of concern?" 13% percent of the respondents from the Nelson House lending center listed the 'standard of customer service' – the only category that had more respondents was 'career prospects.' The employees' concern for service quality exists despite the upward trend in all the CSI global indicators of service quality and management's apparent satisfaction with the

130

performance of the lending center. The dissonance between perception of quality between management and employees was confirmed by some of the interviewees when asked how they felt Delivery Strategy was working for the bank.

> ... the feedback you get back [from the customer] is "I'm dehumanized, I just became a number. I cannot longer talk to you as a person, you just treat me as a number." ... we have lost the customers along the way when we did not clearly communicated DS to them.[26]

> It has its good points. I think with the mass market [personal customers] it seems to work very well. It is not the same thing with small businesses ... they get frustrated, and we get frustrated because we can see them. ... I don't think we are satisfying customers. Well, not small businesses anyway.[27]

> I think it has been very effective in actually improving the lending book because of the monitoring system. Now that we've seen that through [the reduction of the risk index], because of the various pressures on us, we are going to be asked to be more proactive in selling — ultimately, the system is now capable of handling more volume — and that will become very difficult. We just don't have the relationship basis to sell effectively. The customers have said that they become a number; and in a way they have. ... It is difficult to sell that way.[28]

Effect of quality pressure on time per order. Notwithstanding the employees' concern for the standard of customer service, the effects of quality pressure on desired time per order are weak in comparison to the effects of work pressure. To estimate the response of quality pressure on time per order a formulation similar to the one used for work pressure was assumed, and the structure to represent the employee-driven formation of quality standard and quality pressure ($\alpha_{se} = \gamma_e = 1$) was added to the constraints used to estimate the parameters that determine time per order.

$$ t_q = f\left(p_q\right) = Exp\left(\gamma p_q\right) $$

Non-linear least squares estimation was used to determine the coefficient for the effect of quality pressure in time per order (γ) and the initial values for customer service expectations (C_0) and quality standard (S_0). As *a fortiori* assumptions for the conservation of quality pressure, customer quality expectations were assumed constant ($\alpha_{ci} = 1$). Employees' perception of service quality was initialized at the initial level of quality delivered, and the time constants for the employees perception of service quality (τ_{qe}) and the formation of quality standard (τ_{sq}) were set at 4 and 26 weeks respectively.

---

[26] TV 6/28/95
[27] PM 6/29/95
[28] MG 6/29/95

$$\underset{\gamma, C_0, Q_{e0}}{Min} \sum_{t=53}^{104} \left(T(\tilde{t}) - TPO(\tilde{t})\right)^2 \qquad for\{\tilde{t} \mid \tilde{t} \in t \text{ and } TPO(t) = value\}$$

Subject to

$$T(t) = T^*(t) \cdot t_w(t) \cdot t_q(t)$$

$$T^*(t) = T^*(t - dt) + \min\left(0, dt\left((T(t) - T^*(t - dt))/\tau_{to}\right)\right)$$

$$sc^*(t) = OF(t) \cdot T^*(t)$$

$$p_w(t) = \left(sc^*(t) - ETA(t)\right)/ETA(t)$$

$$t_w(t) = Exp\left(\alpha p_w(t)\right)$$

$$q(t) = f^*\left((T - C_o)/C_o\right)$$

$$Q_e(t) = Q_e(t - dt) + dt\left((q(t) - Q_e(t - dt))/\tau_{qe}\right)$$

$$S(t) = S(t - dt) + dt\left((Q_e(t) - S(t - dt))/\tau_{sq}\right)$$

$$p_q(t) = \left(S(t) - Q_e(t)\right)/S(t)$$

$$t_q(t) = Exp\left(\gamma p_q(t)\right)$$

$$\tau_{to} = 18.73 \quad \tau_{qe} = 4 \quad \tau_{sq} = 26 \quad \alpha = -0.64 \quad T_0^* = 1.075$$

$$f^* = function \; in \; Figure \; 3.9$$

The results of the estimation process gave reasonable values and tight confidence intervals for the level of customer expectations and the initial quality standards. However, the effect of quality pressure on time per order was not significantly different from zero (see confidence interval).

```
-0.1967 <= GAMMA =                          -0.0621 <= 0.1067
 1.0078 <= CUSTOMERS' EXPECTATIONS = 1.1652 <= 1.2428
 0.8693 <= INITIAL QS =                      0.9554 <= 1.0532
```

Furthermore, the introduction of quality pressure into the estimation of time per order does not increase the R^2 of the fit to the historical data significantly – from 0.828 to 0.829.

An alternative way of gauging the impact of quality pressure on time per order is by exploring the residuals of the original estimation of time per order. Given the assumptions that the quality standard is formed exclusively by the employees' perception of service quality, it can be shown that the numerator of quality pressure is a function of

132

the *changes* in time per order[29]. If quality pressure had a significant impact on time per order it would be expected that the residuals of a regression without it to be autocorrelated and the summary statistic for historical fit to show a large element of variation. The results of the regression with just the effects of work pressure – besides having a good R^2 (0.82) – do not show any of these characteristics.

The weak response to quality pressure is consistent with the emphasis employees place on processing customer orders in the same day they arrive, and the tendency to 'cut corners' as a way to deal with heavy caseloads. Of the 15 loan offices interviewed, all but one mentioned the reduction of efforts to sell additional products and document customer transactions in times of high work pressure. Figure 4.20 shows the results of a regression between the effective time per order – as estimated from the service delivery sector – and the business loans sales (£/week)[30]. Although the sales data have a large variance because of the discrete amounts processed in each transaction, the effective time per order is a significant predictor of the overall behavior of sales, thus, establishing the predicted link between time per order and financial performance. Sales lost due to high work pressure are just one of the hidden costs of 'poor quality.'

---

[29] The transfer function for quality pressure is given by: $P_q(s) = \dfrac{s\tau_{sq}}{\left(s\tau_{qe}+1\right)\left(s\tau_{sq}+1\right)}$.

Doing the inverse Laplace, it is easy to see that the transfer function is the difference of two exponential smoothings of the same input, thus being responsive to changes in time per order but returning to zero if the input does not change; the structure does not show steady state error.

$$p_q(t) = \frac{\tau_{sq}e^{-t/\tau_{qe}} - \tau_{qe}e^{-t/\tau_{sq}}}{\tau_{qe}\left(\tau_{qe}-\tau_{sq}\right)}.$$

[30] The series for business loans sales had to be transformed by its square root to maintain the equality of variance assumption.

SQRT(B. loan sales)	n = 48	F-Ratio = 13.581
R^2 = 0.23	Coefficient	t-stat
Constant	-711.517	-3.1875
Effective TPO	773.666	3.6852
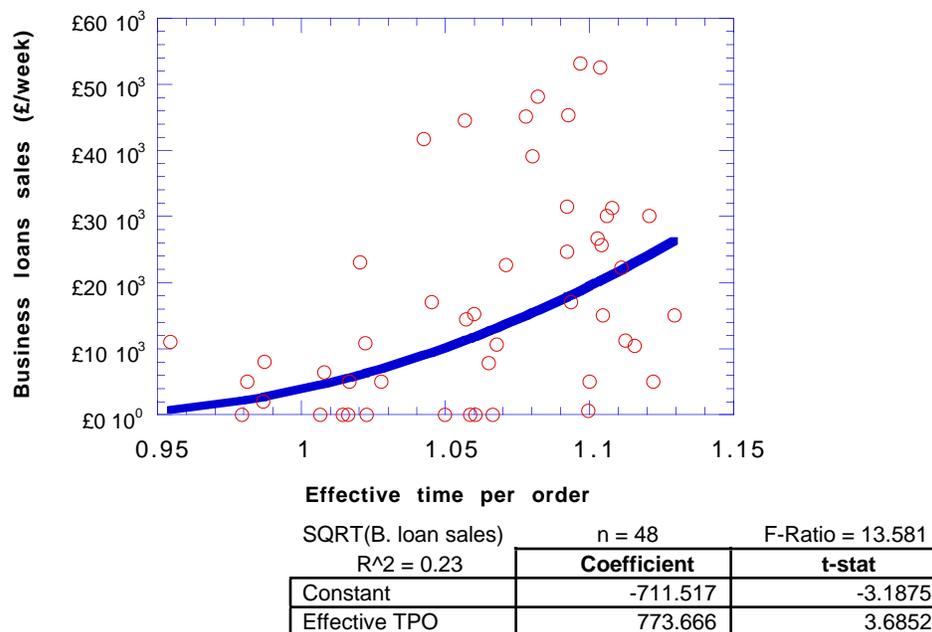
Figure 4.20   Effective time per order as predictor of business loans sales

§4.5. Full System Tests

The overall adequacy of the proposed theory was tested through the model's ability to replicate the historical behavior of the lending center. Additionally, a set of sensitivity tests was performed to determine the significance of the different elements of the theory in the center's operations and the implications of the center's policies were explored through extended simulations under equilibrium conditions.

§4.5.1. Historical Fit of the Model

To test the historical fit of the proposed theory the model was simulated with two exogenous data series driving it: the weekly demand on the lending center (orders/week) and the weekly rate of absenteeism (employees/week). Both of these series have a significant random component and are outside the model boundary[31]. All model parameters were set to the values estimated in §4.4, and the model was initialized to reflect the status of the lending center at the beginning of the second year of operations (week 52). Table 4.3 lists the system parameters and initial conditions used in the simulation.

---

[31] Because of the relatively low work intensity throughout the period were data were available, no significant endogenous explanations for the absenteeism rate were found.

Service Capacity

Labor

τ_l	Time to adjust labor	8.89	week
λ_l	Hiring delay	8.98	week
τ_v	Time to cancel vacancies	1.0	week
τ_t	Time for turnover	401	week
τ_{ld}	Time to implement labor layoffs	1e99	week
χ	Maximum labor growth rate	0.06	Frac.
τ_e	Time for experience	12.0	week
ε	Relative effectiveness of rookies	0.35	Frac.
η	Fraction of experienced personnel for training	0.05	Frac.

Capital

τ_k	Time to adjust capital	8.89	week
λ_k	Capital acquisition delay	6.0	week
τ_s	Time to cancel capital on order	1.0	week
τ_d	Average capital life	520	week
τ_{kd}	Time to implement capital sales	1e99	week
φ	Max. capital growth rate	0.06	Frac.
ϕ	Capital technological content	35.0	hour/week
θ	Capital labor intensity	1.0	emp/cap

Factor Demand

τ_{pe}	Time to perceive labor effectiveness	6.70	week
τ_{l*}	Time to adjust desired labor	18.7	week
σ	Elasticity of substitution of production factors	0.0	
κ	Capital intensity	0.5	Frac.
π_l	Price of Labor	1.0	$/emp/wk
π_k	Price of Capital	1.0	$/cap/wk

Table function slopes

α	Effect of work pressure on time per order	-0.64	
β	Effect of work pressure on work intensity	0.37	
γ	Effect of quality pressure on time per order	0.00	

Service Delivery

τ_r	Minimum residence time for an order	0.1	week
τ_p	Minimum processing time per order	NA	week
λ_σ	Desired delivery delay	0.1	week
τ_{to}	Time to adjust desired time per order	18.7	week
τ_{fp}	Time to accumulate fatigue for effect on productivity	3.0	week
τ_{ft}	Time to accumulate fatigue for effect on turnover	52.0	week

Service Quality

α	Responsiveness coefficients for aspiration adjustment rates (8)	$\alpha_{se}=1$ $\alpha_{ci}=1$	
γ	Responsiveness coefficients for quality pressure (3)	$\gamma_e=1$	
τ_{qc}	Time to identify customers' perceptions of quality	NA	week
τ_{qe}	Time for employees to adjust quality perception	4.0	week
τ_{qm}	Time for managers to adjust quality perception	NA	week
τ_{cq}	Time for customers to adjust service expectations	1e99	week
τ_{sq}	Time to adjust quality standard	NA	week
τ_{gq}	Time to adjust quality goal	NA	week
β_c	Customers' quality perception bias	0.0	quality
β_e	Employees' quality perception bias	0.0	quality
β_m	Management's quality perception bias	0.0	quality
μ	Customers' expectations reference	1.165	hour/order
ψ	Professional quality standard	NA	quality
ξ	Management quality standard	NA	quality

Initial conditions

L_e	Experienced personnel	64.1	emp
L_r	Rookies	14.1	emp
K	Capital	79.9	capital
T^*	Desired time per order	1.07	hour/order
E	Perception of labor effectiveness	0.78	Frac.
Q_s	Quality standard	0.95	quality

NA: Not applicable, has no impact on the simulation results.

Table 4.3   Nelson House LC. System parameters and initial conditions

The summary statistics for the historical fit of the model to the data series available is shown in table 4.4, and figure 4.21 shows the behavior of the simulated data against the historical series.

|			Theil's Inequality Statistics				
	MAPE	MSE	Bias	Unequal Variation	Unequal Covar	R^2	N
Desired labor	0.9%	1.145	0.148	0.278	0.574	0.723	52
Total labor	1.1%	1.012	0.063	0.327	0.609	0.662	52
Time available	1.3%	1262.61	0.067	0.095	0.839	0.898	50
Orders processed	0.3%	82.50	0.000	0.294	0.706	0.990	50
Time per order	1.8%	0.001	0.133	0.041	0.826	0.748	50
Work intensity	1.7%	0.000	0.042	0.159	0.799	0.638	50

Table 4.4   Historical fit of model

The Mean Absolute Percent Error (MAPE) between the simulated and actual variables is less than 2% for all series, indicating a close fit of the model to the actual behavior of the lending center, and the low bias and variation components of the Theil inequality statistics indicate that the errors are unsystematic.

The model's good tracking of the historical series of orders processed is due to the fact that the series in only one step removed from the customer orders data series. The relatively low R^2 in some of the comparisons are caused by the small trend component that most of the data series have and the high frequency variations of the driving data series. The estimated model is functioning as a low-pass filter capable of tracking the overall behavior of the system variables but not very reliable for point predictions of high frequency events.

Matching historical behavior only tests the replicative validity of a model. A full test of model's representativeness has also to consider its structural validity (face validity). The derivation of the model structure and parameters from observed micro-decisions and physical flows in the LC presented in §4.4, and the ability for partial model structure to replicate intermediate data series constitute true tests of the model's structural validity.

Having ascertained representative validity, the next section comments on the observed and estimated policies from the LC, their adherence to the relationships hypothesized in the proposed theory of service delivery and their significance on the overall behavior of the system.
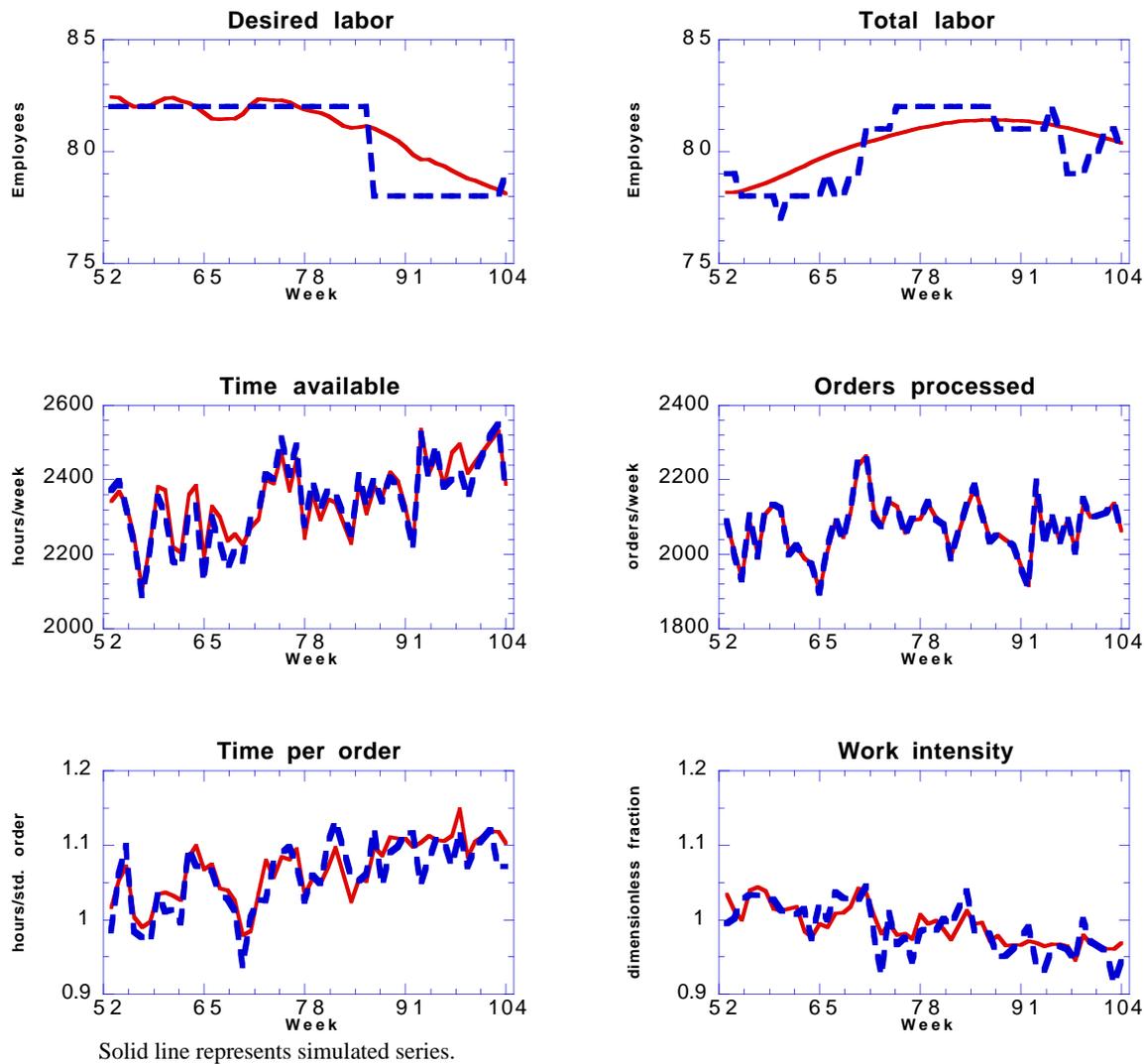
Figure 4.21   Historical fit of model

§4.5.2. Significance of Behavioral Components

For presentation purposes, the proposed theory of service delivery has been broken down
into four behavioral components:

i) Managerial hiring policies respond only to throughput goals and do not consider the
real productivity of service personnel; management does not perceive the skill mix
and the effects of fatigue on productivity.

ii) Recently hired employees have to go through a training period before they become
fully productive.

iii) Employees respond to changes in work pressure by modifying the time to process
an order and, to a lesser degree, their work intensity.

iv) The effect of quality pressure in maintaining an adequate time per order is relatively weak because of the lack of metrics to monitor quality.

To test the significance of the different elements of the proposed theory in explaining the quality performance of the Nelson House LC, a set of simulations was performed varying system parameters that affect each of these elements. The simulations were compared using cumulative delivered quality (*q*) as the main indicator of performance. The simulation to match historical performance was used as the base simulation, and all simulations had the same initial conditions and duration.

Table 4.5 shows the percentage change in cumulative delivered quality as each of the selected parameters was varied ±25%. To test the sensibility of the system to each parameter, only one parameter was modified per simulation. The results of these tests have to be interpreted cautiously because during the second year of operations the LC was still in transient behavior after the buildup stage (consider the gap between desired labor and total labor in figure 4.21). The tests, however, can be used to detect policies with greater leverage.

PARAMETER	Base value		%Δ in cumulative delivered quality	
			Base value * 0.75	Base value * 1.25
Hiring responsiveness				
T to adjust desired labor*	18.76	weeks	-1.9%	1.4%
T to adjust labor*	8.90	weeks	6.0%	-5.4%
Hiring delay*	8.98	weeks	6.1%	-5.4%
T to perceive labor productivity	6.70	weeks	0.7%	-0.6%
Employees' learning curve				
T for experience	12.00	weeks	28.0%	-26.7%
Frac. experienced pers. for training	0.05	fraction	2.8%	-2.9%
Rookies' effectiveness	0.35	fraction	-20.6%	19.1%
T for turnover*	400.78	weeks	-11.8%	6.8%
Response to work pressure				
T to adjust desired T per order	18.73	weeks	-10.6%	6.8%
ALPHA (wp->TPO)*	-0.639	slope	2.3%	-15.4%
BETA (wp->wi)*	0.371	slope	-4.1%	-2.6%

* Denotes that the ±25% interval is wider than the 95% confidence interval of the parameter estimator.

Table 4.5   Impact of parameter changes on quality payoff

Management hiring policies. The proposed model structure was capable of reflecting the staffing policies operating in the learning center and the estimates of the system parameters corresponded with the behavior reported by interviewees. However, contrary

to what was predicted by the theory, management seems to have a relatively accurate perception of labor productivity (short time constant). The close monitoring of labor productivity was explained by the absence of prior productivity figures for the LC. There is evidence from interviews that the adjustment process for the perception of labor productivity is asymmetric (it only adjusted upward), and that the estimation of labor requirements does not consider the learning curve (employees are assumed to have the same productivity). Unfortunately, the data available did not allowed confirmation of these comments.

Regarding the sensitivity to systems parameters, the delays to adjust labor and hire personnel have the expected effect on the overall performance of the service center; the shorter these delays are – the more responsive the hiring policy is – the better quality is delivered by the LC.

The effect of changing time to adjust desired labor, although limited, seems counterintuitive to the premise that reduced response time should yield better performance. An explanation for this unexpected behavior is that the system was simulated over a period with excess capacity, thus delaying the correction of excess capacity increases the quality performance. Finally, the system is not sensitive to changes in the time it takes management to update their expectations of labor productivity. As explained before, the base estimate is artificially low because of the lack of prior productivity figures, thus it is operating close to its ideal value.

Learning curve. Although labor was relatively stable during the simulation period, performance is extremely sensitive to the employees' learning curve. Reducing the time constant to gain experience from 12 to 9 weeks increases the systems performance by 28%. Similarly, hiring personnel with a higher initial effectiveness increases significantly the simulated quality performance. The effect of time for turnover on overall performance seems to be significant; specially considering the artificially extended job tenure caused by the unemployment in the London area.

Because of its relatively small base value, the changes in the fraction of experienced personnel for training does not have a great impact in the system's overall performance. The difference in response to changes to time for experience and fraction of supervision is noteworthy. Perhaps more supervision or additional training prior to deployment of new hires could effectively be used to reduce the time to gain experience.

Response to work pressure. As predicted by the theory, high work pressure triggers a relatively higher reduction in time per order than the increase in work intensity. However, the asymmetric adjustment of the desired time per order was not expected in the original formulation of the theory. It was particularly surprising to find that the desired time per order was not allowed to increase, but this it could also be explained by the lack of prior productivity standards and the need to generate them

The system is relatively sensitive to changes in the time constant that regulates the downward adjustment of the desired time per order. This sensitivity is to be expected because the absence of feedback loops to increase the desired time per order – there is no effect of quality pressure, and there is no upward adjustment for the desired time per order – thus making the system behavior path dependent.

With the exception of the more aggressive adjustment of time per order as a response to work pressure, the changes in the employees' responses to work pressure do not have significant impact on the overall system performance. Again, these results are biased because of the excess in service capacity available throughout the second half of the simulations.

Response to quality pressure. As predicted by the theory, no operational metrics of service quality were identified in the LC, and no effect of quality pressure on time per order was detected. To test the potential impact of quality pressure on the quality performance of the lending center a pair of simulations were performed modifying the slope of the effect of quality pressure on time per order (originally estimated at zero, i.e., no effect) to one and minus one[32]. In both simulations the change in the quality payoff was less that 1% — the quality pressure stayed within the [-0.025, 0.036] range throughout the simulation. The reason is that the formation of quality pressure that is in place in the lending center – exclusively driven by employees' perceptions – is effective only in correcting sudden decreases in quality, but, because of the delays in perceived quality and erosion of quality standard, it is not successful in dealing with long term trends.

From the sensitivity analysis we can infer that having enough effective service capacity – either by adjusting personnel more quickly, accelerating the training process, or retaining experienced personnel – seems to be the main lever for the delivery of quality. If enough

---

[32] Because the initial estimate for the slope coefficient was 0 it was not possible to perform the ±25% test.

capacity is in place, as was the case for the second half of the simulation, the responses that employees have to work pressure are not very significant. Despite the confounding effects of the transient behavior remaining from the buildup stage of the LC, enough evidence was found during the second year of operations to corroborate each of the behavioral components of the proposed theory. The combination of the proposed elements is hypothesized to cause the long term erosion of service quality. The implications of the current policies over a longer horizon will be explored in the next section.

§4.5.3. Extended Simulations

Although the performance of the LC during its second year of operations (the period matched by the model) seems to be adequate, the results are suspicious because the LC was still in transient behavior during that year after the center's buildup phase. To assess the implications of the current policies of the LC under more stable conditions, the simulation horizon of the model was extended for two years beyond the final point where data were available.

Since the two data series driving the model do not show any significant trend component[33], it was possible to capture their main characteristics with a pink noise random number generator capable of reflecting the autocorrelation seen in the data series (Britting, 1973). Table 4.6 shows the main statistics for the two weekly data series.

	Absenteeism	Customer orders
N of cases	50	50
Minimum	0.101	1893.9
Maximum	0.235	2265.1
Mean	0.166	2071.8
Standard Dev.	0.032	78.3

Table 4.6   Statistics of exogenous data series

Assuming an exponential decay of the autocorrelation coefficients, it is possible to estimate the time constant for the smoothing of the pink noise generator. The autocorrelation function (ACF) of the absenteeism data series is almost 0 at lag 6 indicating a time constant of two weeks for the smoothing process – three time constants account for 95% of the adjustment of an exponential process. Although the ACF of

---

[33] Time is not a significant regressor in a model with autoregression component.

customer orders does not have as a clear profile of exponential decay as the absenteeism's ACF, similar logic was used to derive a time constant of 1 week for that series.

The series parameters were fed to a macro to generate pink noise (Richardson and Pugh, 1981)[34], and the generated series were used to extend the real data series from week 104 to week 208. Figures 4.22 and 4.23 show the complete data series used for the extended simulation—data from week 52 to 104 is historical. It is worth noting that the center's customer demand is quite stable; the normalized standard deviation (σ/μ) of the customer orders series is less than 4%.

Customer orders

Figure 4.22   Customer orders (extended data series)

---

[34] The pink noise macro was implemented in Vensim after Richardson and Pugh (1991).

```
:MACRO: pink noise(mean,std dev,tao,dt)
pink noise = INTEG((mean + white - pink noise)/tao, mean)~~|
white noise = std dev * sqrt(24*tao/dt) * (RANDOM 0 1()-0.5)~
        ~ The subtraction of 0.5 to the random number generator is to
          center the distribution around 0|
:END OF MACRO:
customer orders = pink noise(2071.8,78.3,1,TIME STEP)~~|
absenteeism = pink noise(0.166,0.032,2,TIME STEP)~~|
```
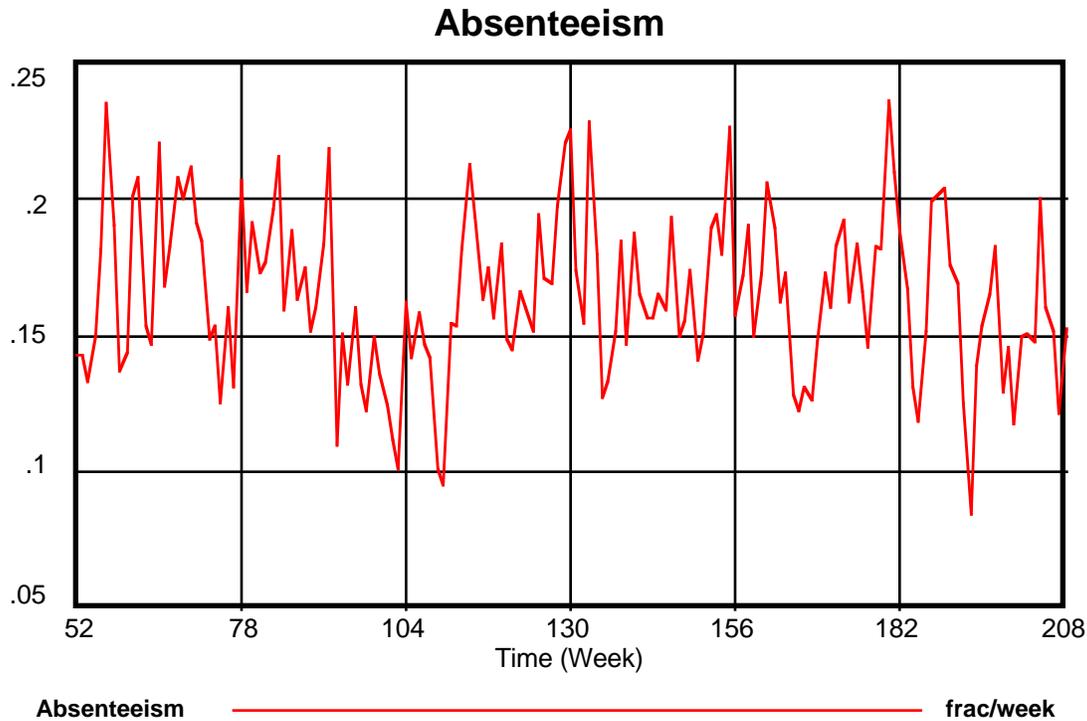
Absenteeism

Figure 4.23   Absenteeism (extended data series)

The model was simulated with the two extended data series initializing it at week 52 and with all the system parameters used in the previous section. The transient from the center's build-up effort can clearly be seen by comparing the simulation output for desired and total labor in Figure 4.24. In this particular simulation it takes until week 146 for total labor to match the desired labor.

Figure 4.25 shows the simulated behavior of desired time per order and the effective time per order. As predicted by the theory, although the system has reached its desired service capacity by week 146 – capacity is in equilibrium – and both of the driving series are stationary, i.e., do not have a trend component, the random variations of customer demand and absenteeism further erode the desired time per order from 1.03 hrs/order in week 146 to 0.99 hrs/order in week 208; an erosion rate of -3.3% per year in equilibrium conditions[35].

---

[35] The erosion rate was calculated using the following formula: $\Delta T = \dfrac{\ln\left(T2/T1\right)}{\left(t_2 - t_1\right)/52}$.

Where T1 represents the desired time per order at the time service capacity is in equilibrium (t_1), and T2 is the last reading if desired time per order available from the simulation (t_2=208). The 52 (weeks/year) factor was used to annualize the erosion rate.
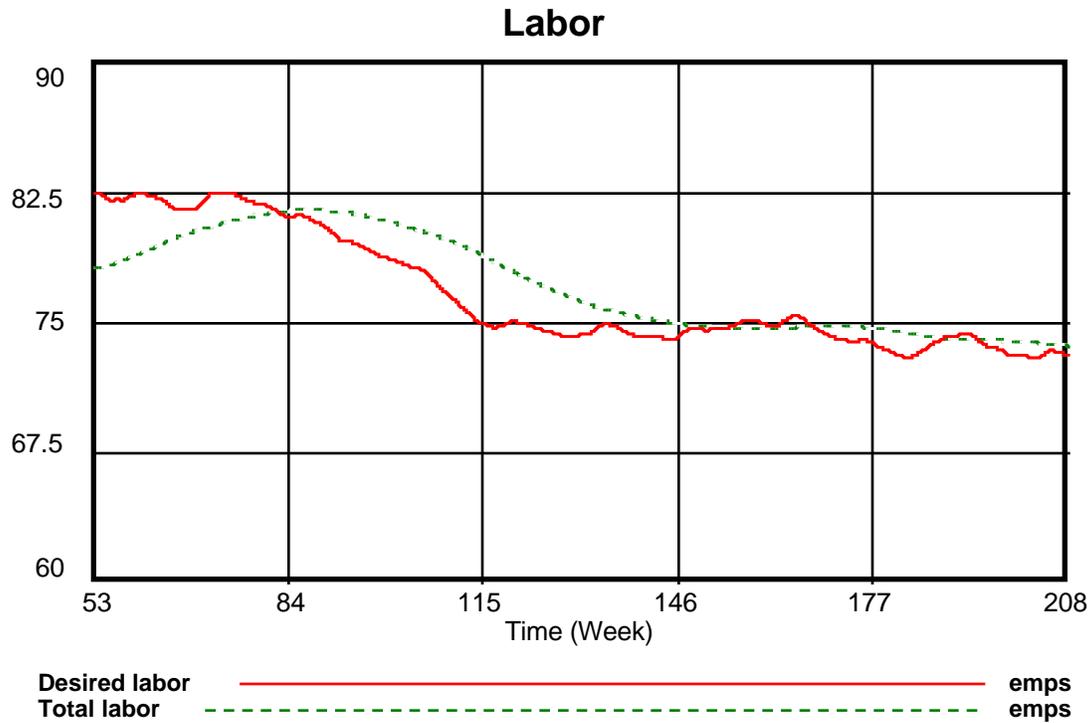
## Labor

Figure 4.24   **Extended simulation. Desired and total labor (simulated data series)**
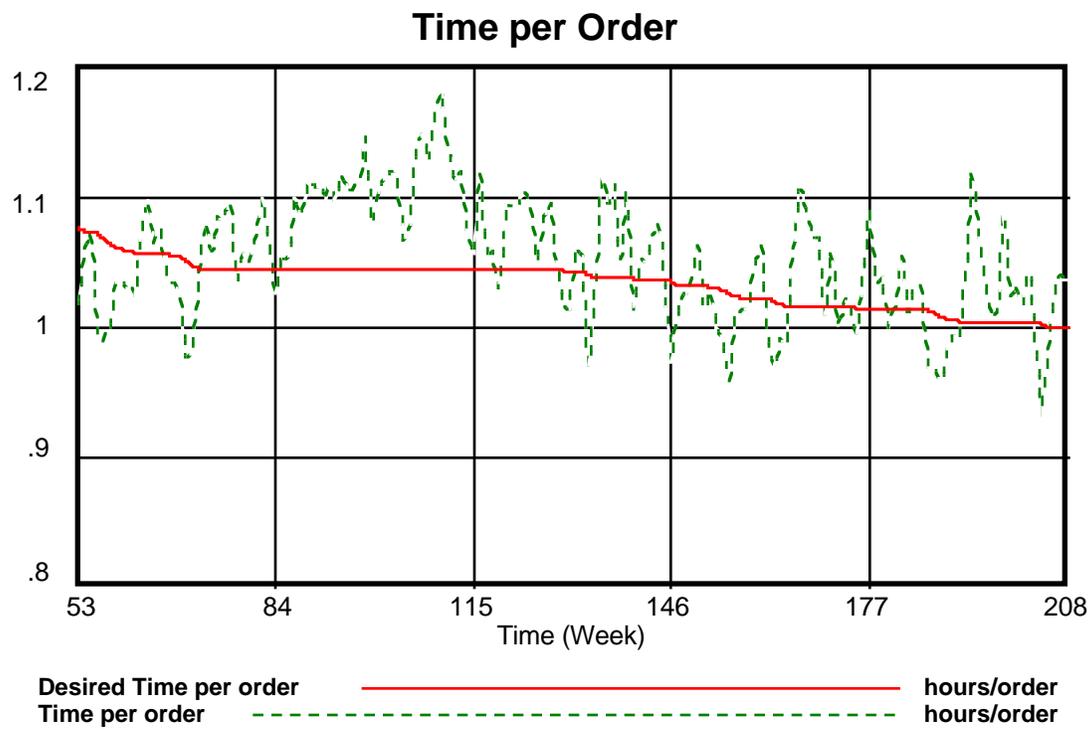
Time per Order

Figure 4.25   **Extended simulation. Desired and actual time per order (simulated data series)**

Because of path dependency, the asymmetric adjustment processes in the model (management perception of labor productivity and the adjustment of the desired time per order) seem to have a significant impact on the historical evolution of desired time per order. Additionally, as commented in the previous section, it was in the asymmetric processes where the calibration process yielded unanticipated results. The impact of these processes in the overall behavior of the system was tested through extended simulations with modified adjustment processes.

The first simulation test was designed to assess the impact of a biased adjustment of the perception of labor productivity as it was suggested by some of the interviewees. The process through which management perceives labor productivity (previously assumed symmetric with a time constant of 6.7 weeks) was made asymmetric with a time constant to update a decrease in productivity three times longer than the time to perceive an increase on productivity; 20.1 and 6.7 weeks respectively.

For the second simulations, the constraint on the upward adjustment of desired time per order was removed to eliminate some of the model's truncated behavior. The time constant to adjust upward was set to be three times longer than the value estimated for the downward adjustment; 56.1 and 18.7 weeks respectively. The last simulation also removes the upward constraint on the adjustment of the desired time per order and makes it symmetrical with an adjustment time constant of 18.7 weeks.

For the three simulations the structural changes to adjustment process were activated at time 104, i.e., the period with historical data was simulated with the original structure. The summary results of the extended simulations, together with the base case, are presented in table 4.7 (all simulations were driven by the same exogenous time series).

Scenario	Capacity in equilibrium		Final reading		Erosion rate
	week	DTPO	week	DTPO	%/year
Base case	146	1.03	208	0.99	-3.32%
Asymmetric adj. of labor prod.	151	1.03	208	0.99	-3.61%
Desired TPO upward adj. (T*3)	133	1.06	208	1.03	-1.99%
Desired TPO symmetric adj.	129	1.09	208	1.07	-0.92%

Table 4.7   Extended simulations. Erosion of time per order

The asymmetric adjustment of the perception of labor productivity fails to recognize expeditiously the effects of absenteeism and experience mix on labor productivity. Hence, management underestimates the labor requirements. Although total labor takes

longer to match the desired labor – there is excess capacity from week 84 until equilibrium is reached at week 151 – the underinvestment in service capacity erodes the desired time per order goal, under equilibrium conditions, at a faster rate than the base case. Allowing for the upward adjustment of desired time per order improves the overall performance of the system, but not enough to stop the erosion of desired time per order once the equilibrium in service capacity is reached[36].

§4.6. Conclusions

This chapter has presented a study where a pre-existing model articulating a theory was modified and calibrated to match the structure and behavior of a service center. Overall, the model – calibrated with information about micro-decisions and internal policies in the service center – provided an excellent explanation of the operational flows and the macro-behavior of the main indicators of the research site, thus increasing our confidence on the structural and replicative validity of the model.

Despite confounding effects, enough evidence was found in the research site to corroborate each of the hypothesized relationships and behavioral components and of the proposed theory of service delivery. Finally, extended simulations showed that, as predicted by the theory, the structural elements of the research site – policies and physical flows – bias its performance towards an erosion of time per order, even when demand and labor force are stable. The managerial implications of these results will be explored in the following chapter.

---

[36] The higher level of desired time per order in the simulations with the upward adjustment of desired time per order is due to the excess capacity originally available in the service center.

5. Managerial Implications

§5.1. Introduction

This chapter reflects back to the managerial world the lessons and implications derived from the proposed theory of service delivery and the validation exercise. The following section presents insights about high-contact services derived from the theory's empirical validation in the Nelson House Lending Center. Section 5.3 derives policy recommendations for managing the erosion of quality in high-contact services. Finally, to facilitate the generalization and transferability of insights, the model is taken outside the high-contact service context, and its usefulness in other service settings is explored. The chapter concludes by summarizing the findings and identifying future research directions.

§5.2. Insights from the Empirical Validation

The process of confronting a theory with the piece of the world that it is attempting to describe normally generates a series of surprising results that constitute an opportunity for updating the theory or making further inferences about the phenomenon being described (Checkland, 1985). This section reflects on two of such findings from the validation effort described in the previous chapter: the relative strength of the different responses to work pressure in the Lending Center, and the formation of the time-per-order objective. The insights generated from the validation effort are included among the managerial implications of this research because they provide managers with a new way to make sense of behaviors observed in service settings.

§5.2.1. On the Response to Work Pressure

Senge and Sterman (Senge and Sterman, 1992) identified three potential responses to high work pressure in a service setting such as the Hanover Insurance case: 1) reduce the time allocated to process each claim, 2) increase work intensity, and 3) expand service capacity. They also identified the time delays, costs, and side-effects involved in the implementation of each mechanism for controlling order throughput. Figure 5.1 presents a simplified causal loop diagram with the three controlling mechanism for throughput (italics [B1-B3]) and their long term side-effects on the center's performance (dark arrows [R1-R3]).
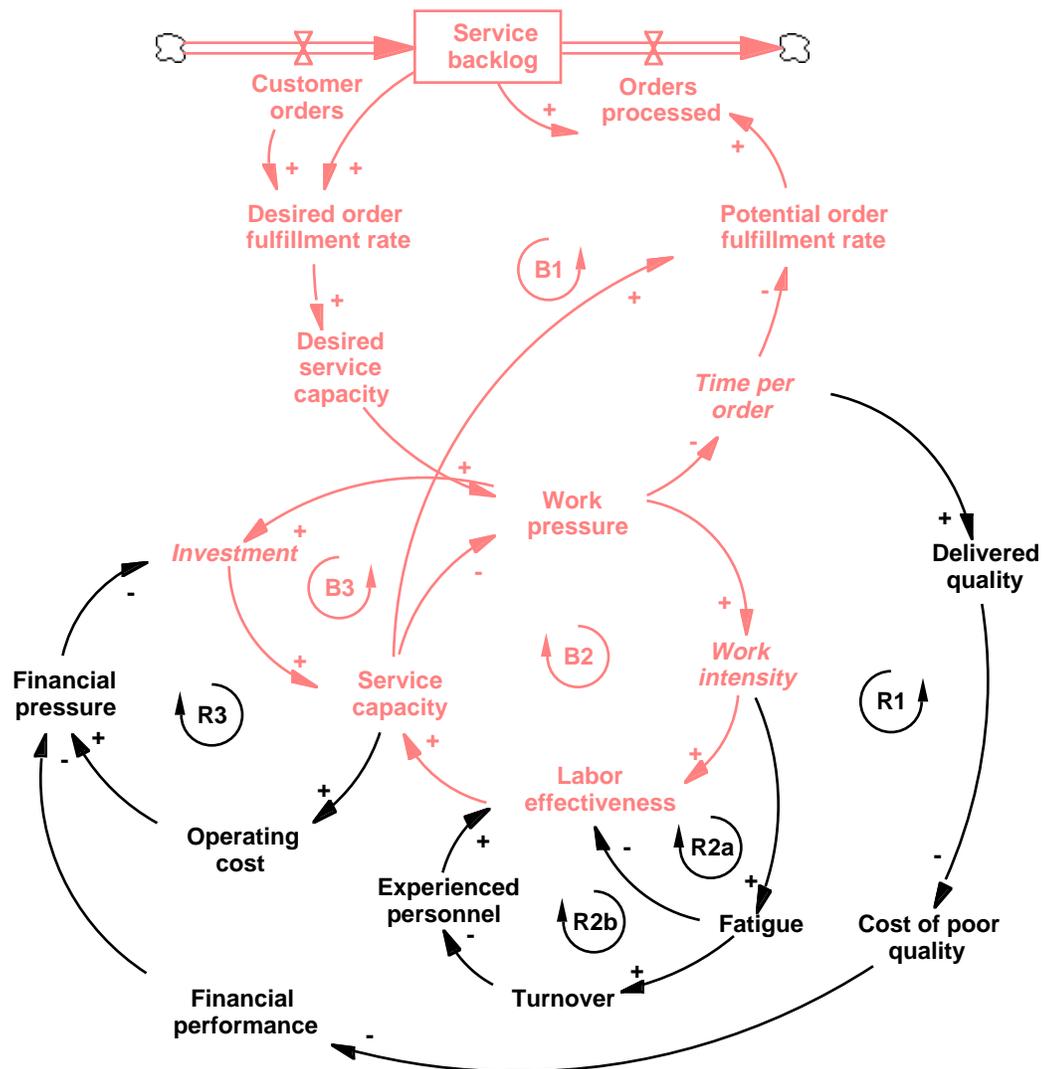
Figure 5.1   CLD. Responses to work pressure and long-term effects

Table 5.1 summarizes the strength (1) and speed (2) of each of these responses as they were estimated for the Nelson House Learning Center. Additionally, table 5.1 also

includes the estimated time constants for the side-effects of the response to have an impact on the performance of the lending center (3), and the time constant of the process through which management perceives this side-effect (4). For example, work intensity shows an elasticity to changes in desired order fulfillment rate of 0.37. Since employees adjust their work intensity as soon as they perceive changes in work pressure, the speed of the response is considered to be instantaneous. Sustained work intensity, however, has the side-effect of reducing labor productivity because of burnout and increased turnover. The time constant for those side-effects to have an impact on productivity was estimated in the LC at three weeks (τ_{fp}). Finally it takes management almost seven weeks to perceive those effects on productivity and start adjusting accordingly (τ_{pe}). The footnotes of table 5.1 explain how the effects of the other responses were calculated.

Response Mechanism	Response		Side Effect	
	Strength (1)	Time for response (2)	Time for impact (3)	Time to perceive (4)
Δ t per order	-0.64	0.0	18.7	∞
Δ work intensity	0.37	0.0	3.0	6.7
Δ service capacity	0.31	18.8+8.9+9.0	0.0	0.0

Table 5.1   Nelson House LC. Summary of responses to work pressure

(1)   Calculated as the elasticity of each response to changes in desired order fulfillment rate at equilibrium. The change in service capacity was estimated as the net increase of productivity as a result of correcting towards desired service capacity $(\varepsilon-\eta)/e$.

(2)   Time constant of the adjustment process that regulates the response to modify throughput. 0.0 means that the adjustment is instantaneous. The hiring response includes the time constant of three successive first order delays: time to adjust desired labor, time to adjust labor, and the hiring delay.

(3)   Time constant of the adjustment process that regulates the long term effects of the response. For the time-per-order response, the time for the long term side-effects is the time to adjust desired time-per-order (τ_{to}); for the work intensity response, the time to accumulate fatigue for effect on productivity (τ_{fp}). The long term effect of service capacity is instantaneous once the extra capacity arrives.

(4)   Time constant of the process to perceive the long term effects of the response. The effects of eroding time-per-order are not detected in the LC because of the lack of operational quality metrics. Employees' fatigue is perceived through labor effectiveness (τ_{pe}). The long term effect of extra service capacity is perceived immediately through the financial system (wages and capital investment).

Figure 5.2a shows the relative strength through time of the three responses to work intensity as a result of a 10% increase in customer orders from full equilibrium conditions. Figure 5.2b shows the integrated effect of the system's reaction for the full simulation[1].

---

[1] These simulations and estimations were made under initial equilibrium conditions and without any noise in customer orders or service capacity.

150

Response to work pressure
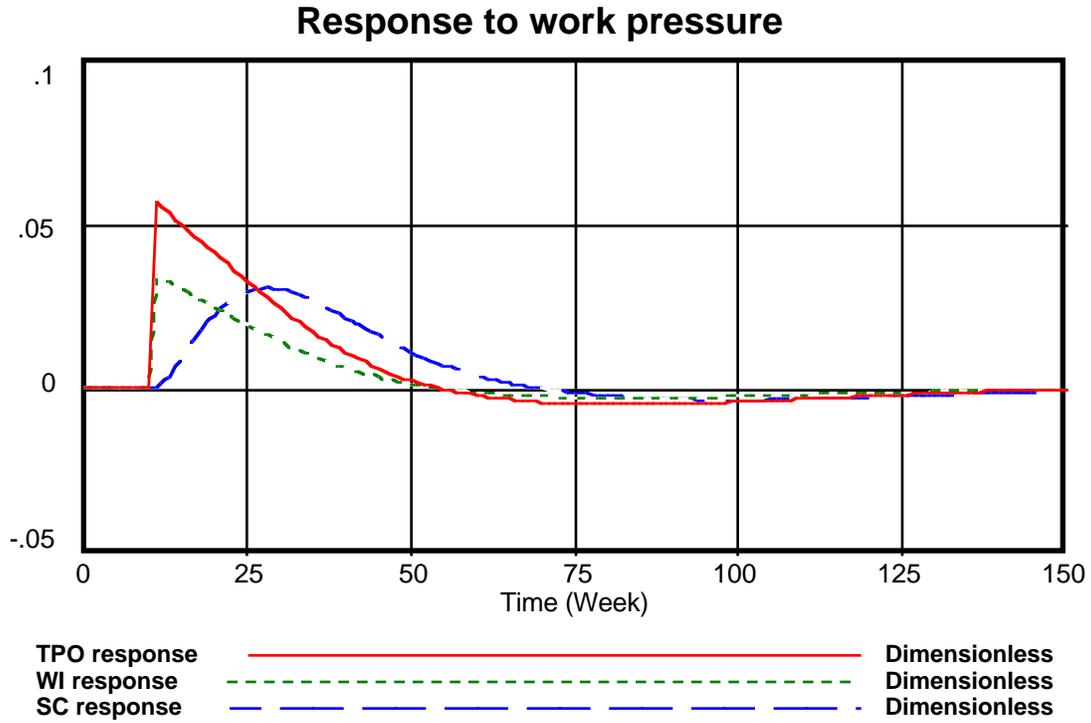
Figure 5.2a   Nelson House LC. Relative responses to work pressure

## Long Term Response (Normalized)
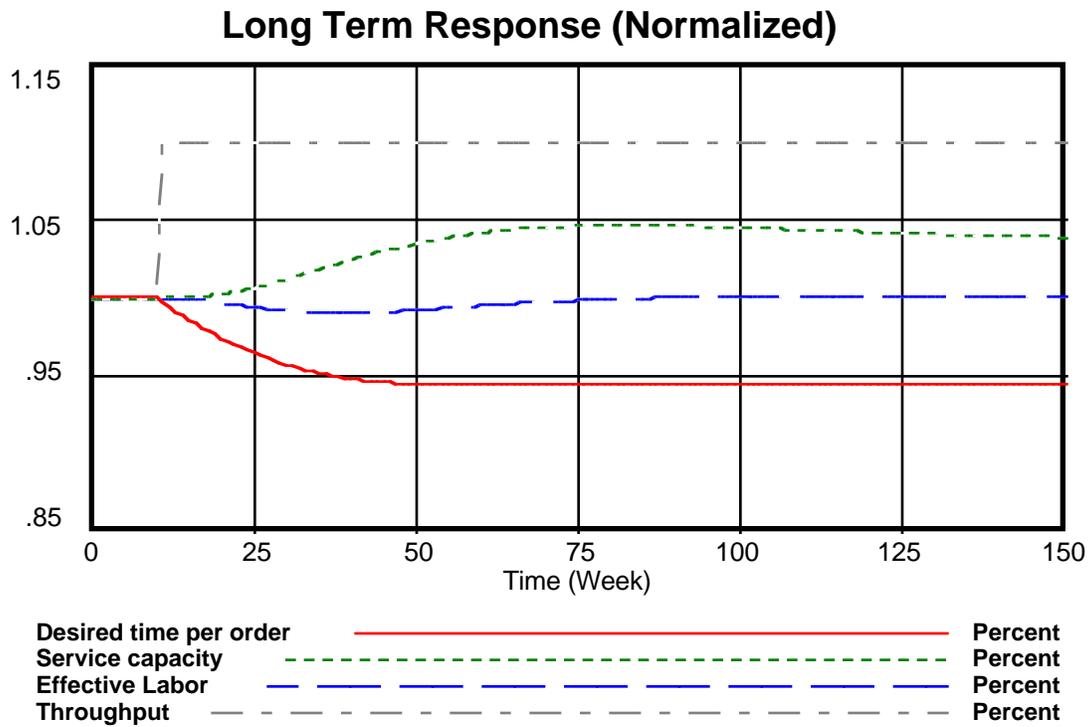
Figure 5.2b   Nelson House LC. Integrated long-term response

The model with the parameters from the Nelson House lending center responded to a 10% increase of customer orders by increasing service capacity 3.9% and reducing of the desired time per order by 5.8%[2].

The first point to note is that the three potential responses to a high work pressure had the same ranking in intensity and speed as the responses observed in the Hanover Insurance case. By analyzing the time it takes for the different responses to have a long term effect in the LC's performance, it is obvious why changes in time-per-order and working intensity are the preferred responses; they provide immediate flexibility without having instantaneous side-effects.

The preference for decreasing time-per-order over work intensity becomes clear when comparing the time it takes each of the responses to have a long-term effect on the performance of the lending center and the time it takes management to perceive that long term effect. While employees start experiencing the effects of fatigue after three weeks delay, the adjustment of the desired time-per-order has a time constant of over four months. Furthermore, management has the mechanisms in place to detect and respond to losses in productivity because of fatigue, but the lack of operational metrics of service quality prevents them from realizing the costs of eroding the quality standard and adjusting service capacity correspondingly.

The structural mechanisms that provide feedback about the long term effects of each of these responses are pervasive throughout high-contact services. All service employees face a loss of productivity from burnout; all services with high customization are susceptible to a reduction in the time to process an order by reducing the service scope and quality. The difference in parameters regulating the impact and perception of side-effects of the various responses is wide enough that it can be safely assumed that, although the values might change from service setting to service setting, the ranking of the relative responses remains the same. Unless management has prior knowledge of these feedback mechanisms, and their implicit delays, there is no other way to gain information about the long-term effects of these responses.

Regardless of the strength of the side-effects caused by each of the responses, the speed of the effect of the responses on throughput (2), and the structural delays and 'fuzzy'

---

[2] The response of the model with the NHLC parameters can be summarized in the following equation:

$$\left(1 + 0.10\right)s_f = \left(1 + 0.039\right)sc\Big/\left(1 - 0.058\right)T^* .$$

information variables in the feedback mechanism to detect its side-effects (4) will always rank the strength of the responses in the order observed in the Hanover Insurance case and in the Nelson House Lending Center. This finding leads to the conjecture that the response pattern might be generic to the high-contact service industry and supports the original hypothesis about the pervasiveness of erosion of service quality in high-contact services.

§5.2.2. On the Formation of Desired Time per Order

One of the most surprising results from the validation exercise was the asymmetry of the adjustment process for the desired time-per-order (T^*). The data from the Nelson House lending center showed that even during periods of low work pressure (p_w), i.e., periods with excess service capacity [between weeks 74 to 104], the underlying desired time-per-order did not adjust upward, although the effective time-per-order (T) was higher than the desired value. From LC's data and interviews, it can be argued that desired time-per-order is eroded by high work pressure, but low work pressure does not have an upward effect. It seems that once employees learn how to deliver the service faster, that ability and mind-set remain with them for future times of high work pressure. If work pressure is reduced, employees will provide a level of service appropriate to the time available. However, the underlying standard of what is achievable will not change.

Intuitively, it is difficult to accept a goal-setting mechanism that does not have a balancing adjustment process to check the erosion of service quality. Quality pressure (p_q) was hypothesized to function as the balancing process to keep service quality on check. However, no effects of quality pressure were detected in the LC and there were no mechanisms in place to introduce quality pressure into the service setting.

On the basis of these two findings (the lack of upward adjustment from work pressure and the absence of an effect of quality pressure), a new conceptualization of the formation of desired time-per-order was developed by separating the effects of quality pressure (t_q) and work pressure (t_w) on time-per-order.

In the revised formulation, the effective time-per-order (T) – the time allocated to process each order – is determined exclusively by the underlying desired time-per-order (T^*) and the effects of work pressure (t_w). If the time allocated to process orders falls below the desired goal for time-per-order, the goal will be eroded adjusting to past performance through an exponential process (d_e).

The effect of quality pressure (t_q), instead of having an immediate impact on the time-per-order, is now assumed to determine an indicated time-per-order ($T^\dagger$) – the time that, according to the quality pressure, should be allocated to process each order. The desired time-per-order (T^*) is adjusted towards the indicated time-per-order ($T^\dagger$) through an exponential process (d_a). Figure 5.2 shows the stock and flow diagram and the corresponding equations of the revised formulation for the formation of the time-per-order goal. The proposed equations substitute equations 41 to 44 in §3.5.

$$(d/dt)T^* = d_a - d_e$$

$$d_a = \left(T^\dagger - T^*\right)/\tau_{ato}$$

$$T^\dagger = t_q \cdot T^*$$

$$t_q = f\left(p_q\right)$$

$$d_e = \max\left(0,\left(T^* - T\right)/\tau_{eto}\right)$$

$$T = \max\left(t_w \cdot T^*, \tau_p\right)$$

$$t_w = f\left(p_w\right)$$
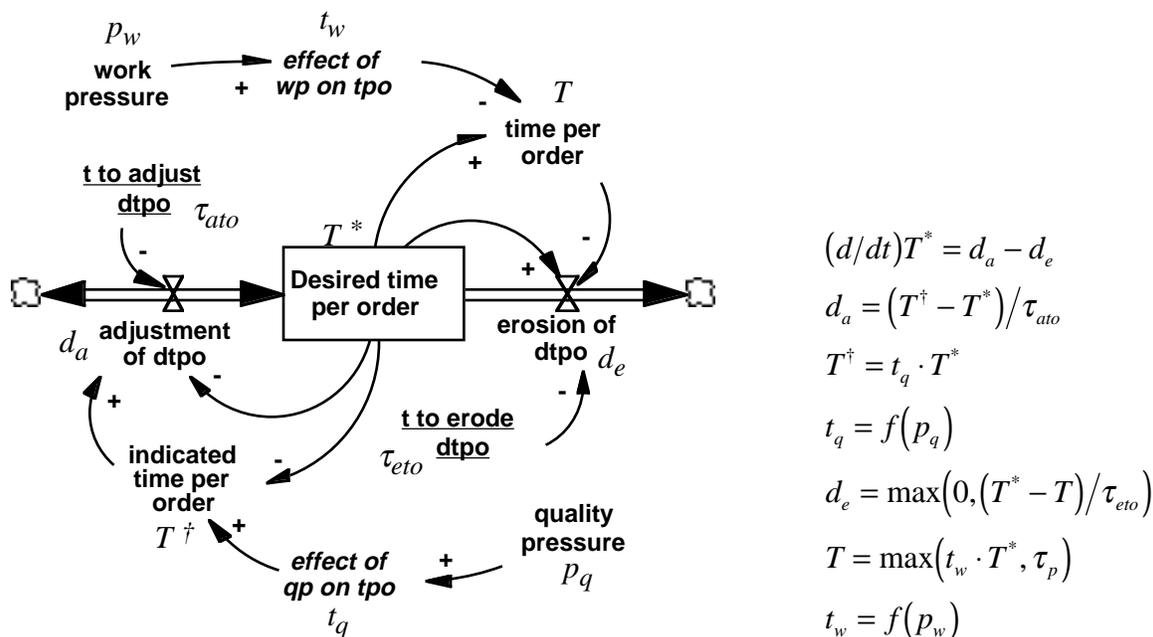
Figure 5.3   Formation of desired time per order. Revised formulation

Although, the absence of upward adjustment of desired time-per-order and quality pressure in the LC does not constitute proof of the proposed formulation, the new formulation has several advantages over the original conceptualization. First, the revised formulation is consistent with the evidence from interviews indicating that quality concerns take a secondary role when employees face work pressure, and the longitudinal data that showed higher time-per-order than desired, but no upward adjustment of the goal[3].

The revised formulation also has the advantage of having a more parsimonious representation of time-per-order (dependent only of the effects of work pressure) and a use of the effect of quality pressure that is consistent with the findings from behavioral

---

[3] The revised formulation is mathematically equivalent to the formulation estimated in §4.4.

decision theory. Finally, the interpretation of the effect of past performance as an erosion mechanism makes for a more compelling conceptualization than the asymmetric formation of aspirations proposed in the original formulation.

Separating the effects of quality pressure and work pressure in the formation of the goal for time-per-order is consistent with other perceptual processes described in the literature that separate motivation and hygiene factors (Herzberg, 1966; Kano, Seraku et al., 1984).

§5.3. Policy Recommendations

As seen in the previous section, the response mechanisms to work pressure ensure that the service setting adjusts to changes in customer demand. The long-term challenge for high-contact service settings is to deliver service at a sustained quality level. This section will derive policy recommendations to maintain service quality while maintaining responsiveness to market demand.

The exploration of policy recommendations will be restricted to those policies possible within the current structure of the model – physical and information flows. Although better policies could perhaps be developed by modifying the content of some information flows, this would entail working with a model for which we no longer have confidence in its structural and representational validity. Because the current structure limits the creation of policy recommendations, this section should be seen as an exploration of a restricted space of possible solutions.

Sensitivity analysis from the previous chapter has shown that the main lever to stop the erosion of time-per-order – the internal proxy for service quality – is to eliminate work pressure, i.e., have adequate service capacity (see §4.5.2). However, the data from the LC have shown that removing work pressure is not enough to upgrade the desired service level.

The discussion in the previous section points to the fact that both maintaining quality pressure and removing work pressure are necessary conditions to avoid the erosion of desired time-per-order, but neither one is sufficient by itself. High quality pressure will always be overridden by the pressures to increase throughput if there is high work pressure, i.e., the effect of work pressure will erode the desired time-per-order at a faster rate (d_e) than the adjustment process (d_a) is capable of compensating for. On the other hand, excess capacity (low work pressure) only stops the erosion rate, but it does not reverse the decline in the desired time-per-order.

The equilibrium conditions for desired time-per-order in the revised formulation (see equation 5.1), show the need of managing simultaneously the effects of work pressure and quality pressure on desired time-per-order.

(5.1)
$$(d/dt)T^* = 0 \Rightarrow \begin{cases} \dfrac{(t_q - 1)}{\tau_{ato}} = \dfrac{(1 - t_w)}{\tau_{eto}} & for\ p_w > 0 \\ p_q = 0 & for\ p_w \leq 0 \end{cases}$$

The following subsections explore the range of possible policies for each of these levers.

§5.3.1. Managing Work Pressure

In §5.2.1 I described the main feedback mechanisms to respond to changes in work pressure – changes in time per order, work intensity, or service capacity – and the long term side-effects associated with each of these responses. In developing a policy recommendation to ensure the delivery of service quality I will aim to minimize the long term effects of the response reducing time per order, i.e., the erosion of the time-per-order goal.

The rate of erosion of the desired time-per-order as a function of work pressure (the normalized gap between desired and actual service capacity) is defined in equation 5.2. It is possible to derive policy recommendations from inspection of the system parameters that control the behavior of the erosion rate.

(5.2)
$$(d/dt)T_e^* = \max\left(0, T^* \left(1 - e^{\alpha(sc^* - sc)/sc} \right) \Big/ \tau_{eto} \right)$$

One possibility to stop the erosion of desired time-per-order is to eliminate the effects of work pressure by isolating the customer serving personnel from any perception of backlog or required throughput. A standard throughput per employee could be established and monitored constantly. Although eliminating variations in the expectation of throughput would ensure a consistent allocation of time-per-order, it would also reduce the center's flexibility to absorb the random variations in customer orders and availability of service capacity[4]. By not giving indicators of work pressure to customer-facing personnel, we would also eliminate the possibility of an immediate response through modifying work intensity and time-per-order. To maintain a reasonable delivery delay,

---

[4] Both sources of variation are immediately reflected in work pressure: $p_w = (sc^* - sc)/sc$.

management would have to carry excess capacity or have access to a flexible reserve capacity to absorb those variations.

An option that would let employees respond to the short term variations in work pressure without incurring in long-term erosion of the underlying desired time-per-order is to reduce the flexibility of the service delivery process (increasing τ_{eto}). The desired time-per-order could be made less flexible by standardizing and/or documenting the service delivery process. Although service personnel could be allowed to 'cut corners' during times of high work pressure, the well-documented process would return to be the service standard as the work pressure is removed from the system. Standardized service delivery processes are, by definition, not appropriate for high-contact services requiring customization, but guidelines and checkpoints for the service interaction could be defined for any transaction. Guidelines and checkpoints could also be used to reduce the effect of work pressure on time per order (α) without the necessity of isolating employees from the work pressure signals.

An alternative strategy would be to ensure that service capacity is acquired before the erosion of time-per-order takes place. By inspection of table 5.1 it is possible to see that in the Lending Center the strength of the adjustment of the desired time per order is only 65% of the adjustment to time-per-order, and that the time constant for the adjustment of service capacity is twice as long as the time constant for the erosion of desired time-per-order. Either we could increase the strength of the response, or ensure that the time constant for expanding service capacity is shorter than the time to erode time-per-order. Equation 5.3 is a simplified version of the rate of change of service capacity as a function of changes in desired service capacity[5].

$$(5.3) \qquad (d/dt)sc = (\varepsilon - \eta)\left(\frac{sc^*/E - (L_r + L_e)}{\tau_h}\right) + (1 - (\varepsilon - \eta))\frac{L_r}{\tau_p}$$

Service capacity can be obtained faster either by having a more responsive hiring policy (reducing τ_h) or by accelerating the process through which rookies become effective service providers (reducing τ_p). Despite the advantages of a responsive adjustment of service capacity in times of high work pressure, the same policy could lead to a fast reduction of service capacity in times of low customer demand, thus removing the

---

[5] The hiring process – adjustment of desired labor, time to adjust labor and the hiring delay – have all been reduced to a first-order adjustment process with a time constant τ_h.

opportunity to increase time-per-order. Alternatively, the strength of the service capacity response could be increased by reducing the effects of the learning curve through hiring people with a higher initial effectiveness or that require less supervision (increasing $\varepsilon-\eta$).

Figure 5.4 (base) shows simulated behavior of the desired time per order in a system with the characteristics of the Nelson House LC – same policies and driving series – but initialized in equilibrium. Consistent with the results in the last chapter, the base case simulation shows a significant erosion of the desired time per order under initial equilibrium. The remaining simulations test – under the same conditions – some of the policies suggested above.

Desired time per order

DTPO base	**Hours/Order**
DTPO low alpha	**Hours/Order**
DTPO slow erosion	**Hours/Order**
DTPO quick labor adjustment	**Hours/Order**
DTPO quick training	**Hours/Order**
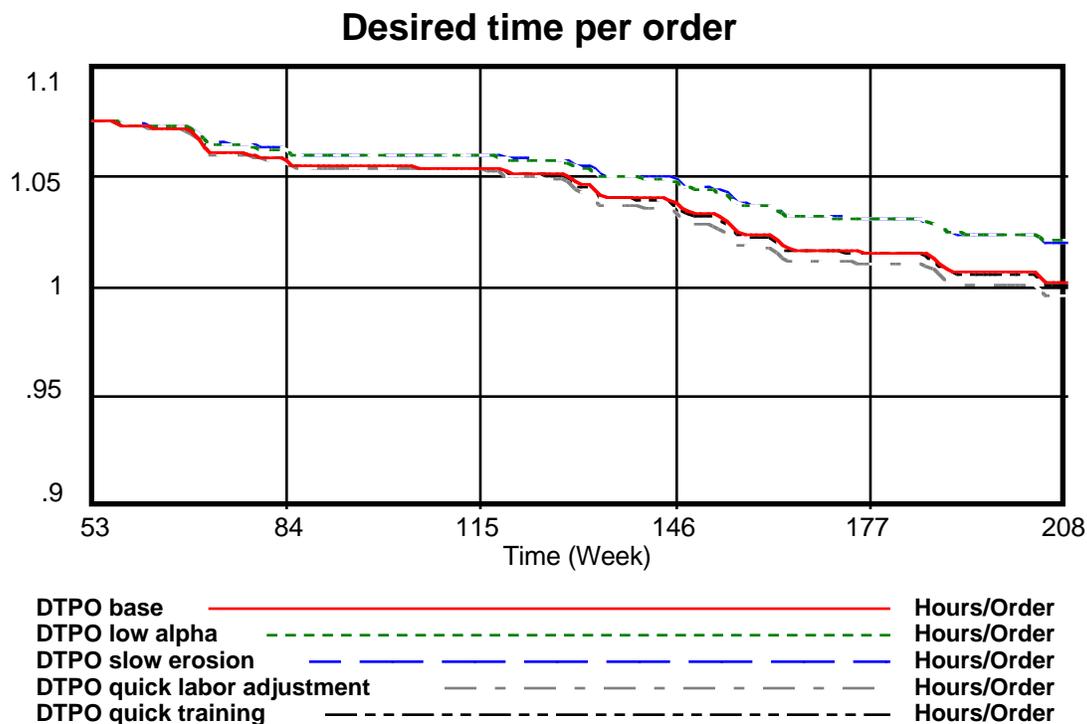
Figure 5.4   Test of policy recommendations. Desired time per order

The 'low alpha' and 'slow erosion' simulations (overlapping) show the effect of reducing the effect of work pressure on time per order (α) by 50% and changing the time constant for the erosion of time-per-order from 18.7 to 28 weeks (50% increase) respectively. Although unrealistic for a high-contact service, this policy has clear leverage towards the sustainability of service quality.

The third simulation (quick labor adjustment) shows the results of reducing the hiring time constant from 36.7 to 25 weeks (Time to adjust labor was reduced 30% to 13 weeks

and the time to adjust labor and hiring delay were reduced 33% to 6 weeks). The higher erosion rate of desired time-per-order under this policy is counterintuitive. One would expect lower erosion rates if service capacity was brought in before the long term side-effects of reduction of time per order could have an impact. In this case, however, the higher erosion of desired time-per-customer is due to the faster reduction of excess capacity during times of low work pressure and the 'worse before better' effect when expanding service capacity.

For the last simulation (quick training – overlapping with the base case) the strength of the service capacity response has been increased by reducing the time required for training by 50% (down to 6 weeks) and increasing the relative effectiveness of rookies from 35% to 50%. The acquisition of service capacity through more efficient training programs has limited effect in the equilibrium situation with very low turnover. Sensitivity analysis in the previous chapter has shown that this policy is a major leverage point in expansion or with higher situations with higher turnover. Table 5.2 summarizes the results of the policy recommendations' simulation tests.

Case	Final DTPO	Erosion rate %/year
base	1.001	-2.39%
low alpha	1.020	-1.76%
slow erosion	1.020	-1.76%
quick labor adjustment	0.996	-2.57%
quick training	1.000	-2.43%

Table 5.2   Test of policy recommendations. Erosion of desired time per order

The intuition to take away from these simulations is that the strength of the service capacity adjustment policies is not driven exclusively by managerial policies. The relative effectiveness of rookies and the time it takes them to become fully effective are also limitations to the responsiveness of this policy. Under conditions of slow adjustment of service capacity the best alternative is to find mechanisms to reduce the erosion rate of desired time-per-order.

Although the suggested policies can reduce the erosion rate of service quality, because of the random variations in work pressure there will be times when it becomes necessary to increase the desired time-per-order. The next subsection explores the possible policies to maintain an upward adjustment process for the desired time per order.

§5.3.2. Maintaining Quality Pressure

The sensitivity analysis results in §4.5.2 showed that even increasing the effect of quality pressure on time per order (initially estimated at zero) has no effect on the overall performance of the LC. The reason for the lack of responsiveness is that the formation of quality pressure that is in place in the lending center – exclusively driven by employees' perception – is effective only in correcting sudden decreases in quality, but, because of the delays in perceived quality and erosion of quality standard, it is not successful in dealing with long term trends. The main challenge is to maintain enough quality pressure to drive the upward adjustment process of desired time-per-order. Figure 5.5 shows the main feedback mechanism to maintain quality pressure[6].
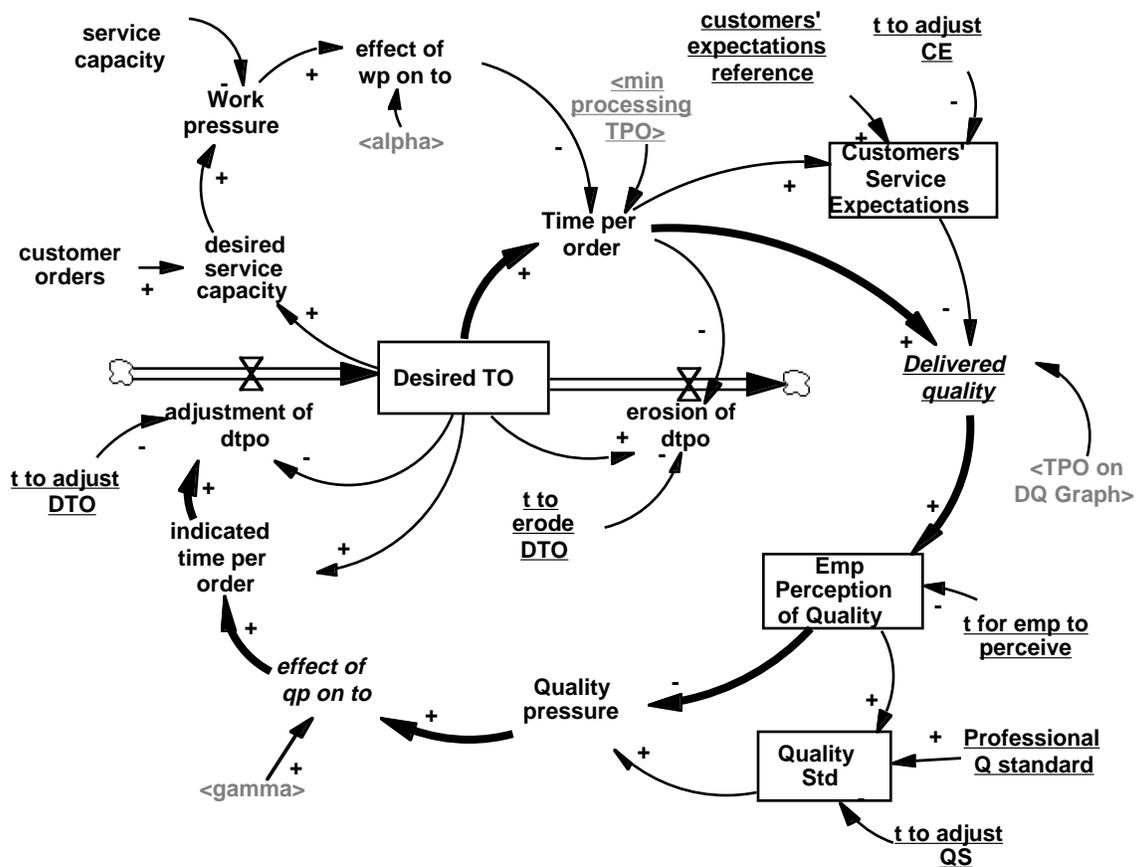
**Figure 5.5   Causal loop diagram. Adjustment of desired time per order**

Several characteristics undermine the efficiency of the controller through quality pressure. First, there are two delays in the information flows – time for employees to

---

[6] For presentation purposes the alternative sources of quality pressure (managerial goal and reports on customer satisfaction) have been omitted.

perceive delivered quality and the time to adjust desired time per order. Second, the non-linear function that determines service quality as a function of the gap between performance and expectations presents a 'flat spot' or 'dead zone' around equilibrium (see figure 3.9) corresponding to a tolerance zone of service delivery (Parasuraman, Zeithaml and Berry, 1994; Strandvik, 1994). The 'flat spot' reduces the responsiveness of the controller to minor changes in performance. Finally, the two aspiration formation processes based on past performance – customers' service expectations and employees' quality standard – erode the long term effectiveness of the controller. Although the erosion of customer expectations does not represent a problem per se, it normally happens at a much slower rate than the erosion of internal quality standards (especially if customers have access to other service suppliers).

Figures 5.6a and 5.6b show the response of the basic controller, i.e., the feedback mechanism without erosion of levels of aspiration, to a 25% step increase in customer orders sustained from week 2 to week 75. The system was initialized in equilibrium ($T^*=C^*=\mu$ and $q= Q_e=S^*=\psi$) with identical slopes for the effects of work pressure and quality pressure ($-\alpha = \gamma = 0.5$) and fixed aspiration levels for customers and employees ($\tau_{cq} = \tau_{sq} = 1e99$). The time for employees to perceive quality was set to 4 weeks, and the time constants for the adjustment and erosion of desired time per order were set to 8 and 16 weeks respectively.

It is interesting to note that while the system remains under work pressure (the first 75 weeks), the feedback mechanism is not capable of bring the desired time per order back to its initial level. The quality feedback mechanism only stops the further erosion of the desired time-per-order, i.e., it equates the adjustment rate to compensate for the erosion rate. The system effectively behaves as a proportional-plus-integral controller responding to the difference between the adjustment and erosion rate (Ogata, 1990)[7]. The only way to increase the desired time per order is by eliminating the work pressure, i.e., eliminating the erosion rate.

---

[7] The reinforcing and balancing loops between desired time per order and its adjustment and erosion rates make the system more complicated. However, its general behavior matches the proportional-plus-integral control action.

Time per order

Customers' service expectations	—————————	hrs/order
Desired TPO	– – – – – – – – – –	hrs/order
Time per Order	— — — — — — — —	hrs/order
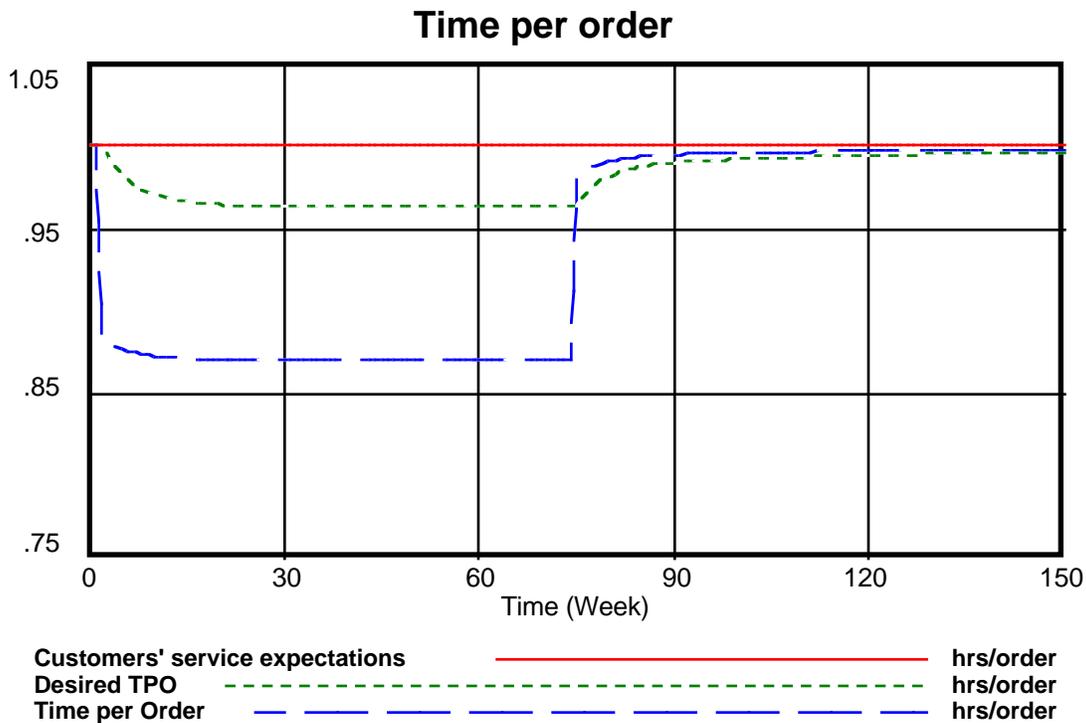
Figure 5.6a   Base case simulation results. Desired time per order

Desired time per order -- rates

| adjustment of dtpo | ————————— | (hrs/order)/week |
| erosion of dtpo | – – – – – – – – – – | (hrs/order)/week |
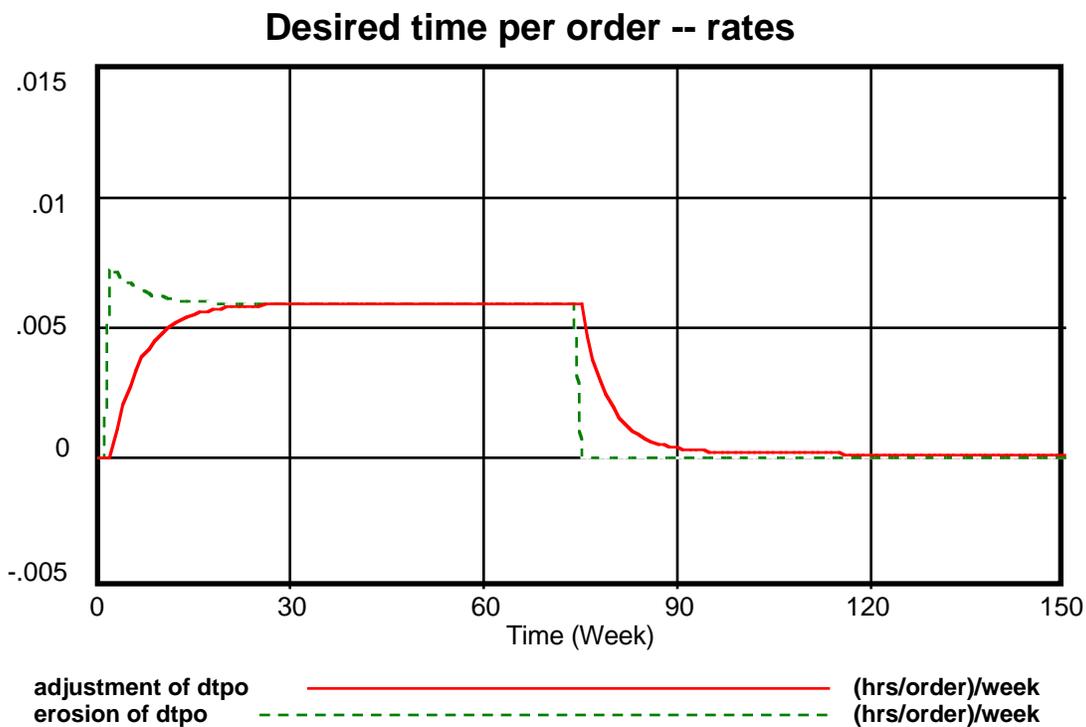
Figure 5.6b   Base case simulation results. Desired time per order – rates

Figures 5.7 show the effect of varying the slope of the effect of quality pressure (the γ coefficient) on the desired time per order and the adjustment rate of the desired time per order. By inspection of equation 5.1, it is possible to see how increasing the effective time per order would increase the equilibrium level of desired time per order when the system is under work pressure. Table 5.3 sumarizes the results of these tests.

gamma	DTPO (week 75)	erosion rate (week 75)	DTPO (week 150)
0.25	0.908	0.0036	0.957
0.50	0.964	0.0058	1.000
0.75	1.000	0.0073	1.000
1.00	1.025	0.0084	1.000

Table 5.3   Sensitivity to gamma. Desired time per order

By increasing the value of the response to quality pressure (γ) we are effectively increasing the open loop gain from delivered quality (q) to the adjustment of desired time per order rate (d_a)[8]. Increasing the gain of the controller increases the aggressiveness of the response to the point where the system overcompensates and reaches equilibrium at a higher quality level ($\gamma=1.0$).

If the internal quality standard is adjustable to past performance, the dynamic behavior of the feedback mechanism changes significantly. Figures 5.8a and 5.8b show the response of the system in which the quality standard is based on past performance with an adjustment time constant of 26 weeks.

An erosion of the quality standard towards past performance reduces the quality pressure felt by employees – the dissonance created by the gap between their perception of delivered quality and what they believe should be delivered to customers. The reduced quality pressure effectively decreases the gain of the feedback response to increase the adjustment rate of desired time-per-order. Consequently, desired time per order continues to erode while there is work pressure in the system.

---

[8] The open loop transfer function from delivered quality to adjustment of desired time per order rate is given by:

$$G(s) = \frac{L[d_a]}{L[q]} = \frac{\gamma}{(s\tau_{ato} + 1)(s\tau_{qe} + 1)}$$

Desired time per order

Desired TPO (gamma=0.25)		Hours/Order
Desired TPO (gamma=0.50)		Hours/Order
Desired TPO (gamma=0.75)		Hours/Order
Desired TPO (gamma=1.00)		Hours/Order
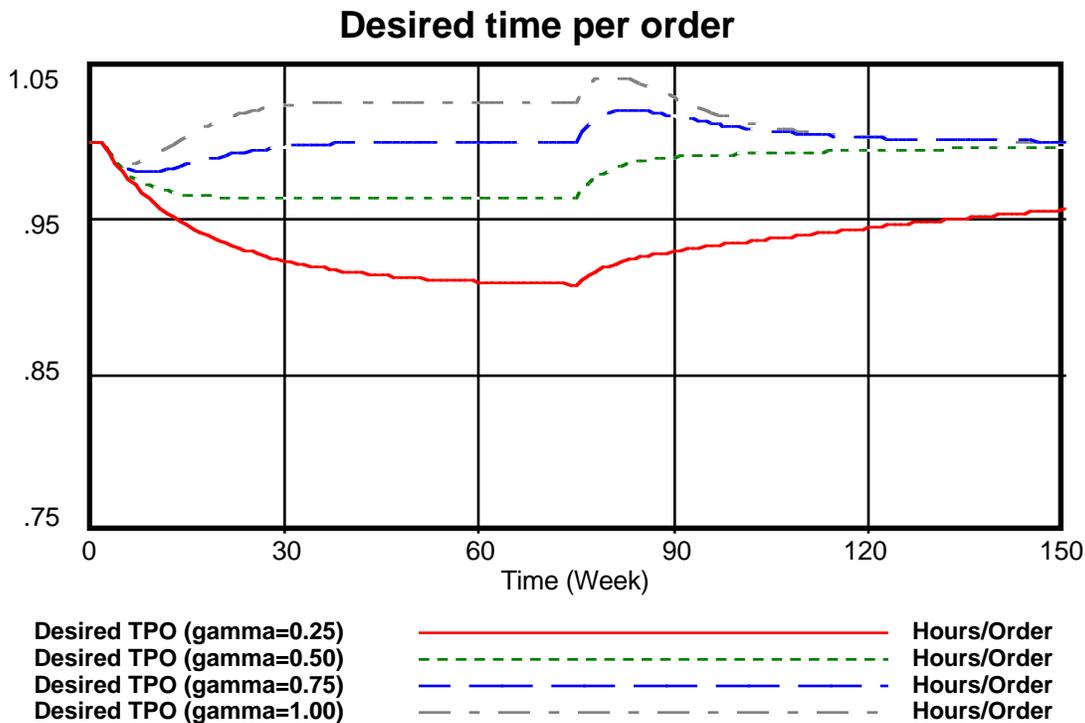
Figure 5.7a   Sensitivity to gamma. Desired time per order

Adjustment of desired time per order

Adjustment of dtpo (gamma=0.25)		(hrs/order)/week
Adjustment of dtpo (gamma=0.50)		(hrs/order)/week
Adjustment of dtpo (gamma=0.75)		(hrs/order)/week
Adjustment of dtpo (gamma=1.00)		(hrs/order)/week
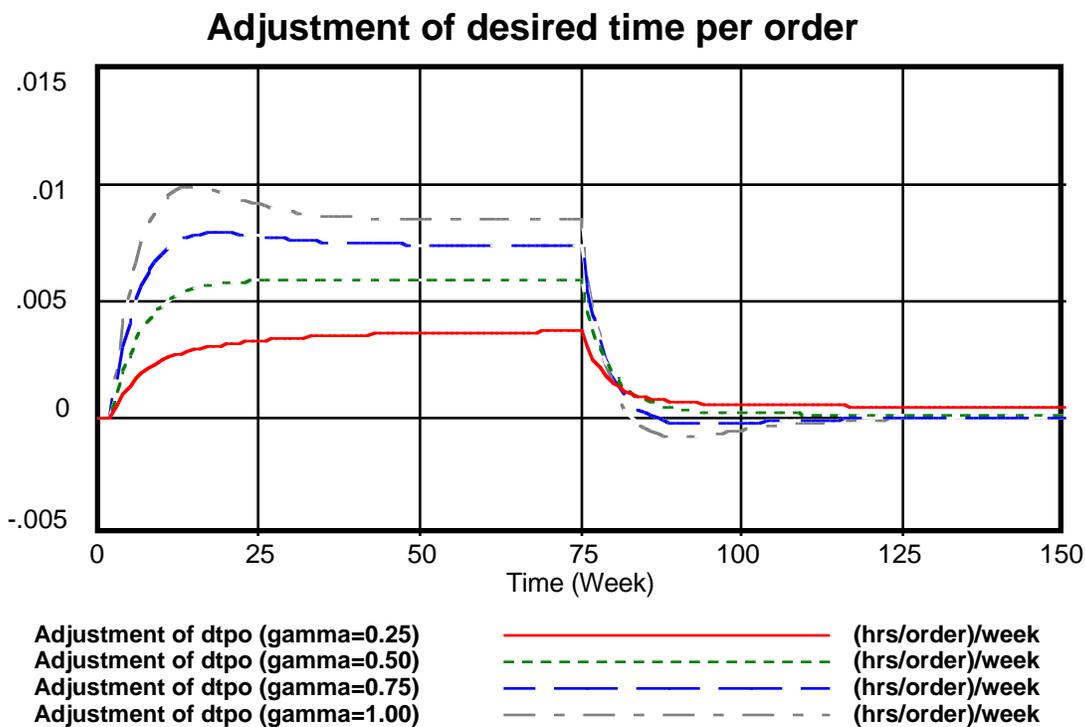
Figure 5.7b   Sensitivity to gamma. Adjustment of desired time per order

164

Time per order

Customers' service expectations	————————————	hrs/order
Desired TPO	- - - - - - - - - - - - - -	hrs/order
Time per Order	— — — — — — — — — —	hrs/order
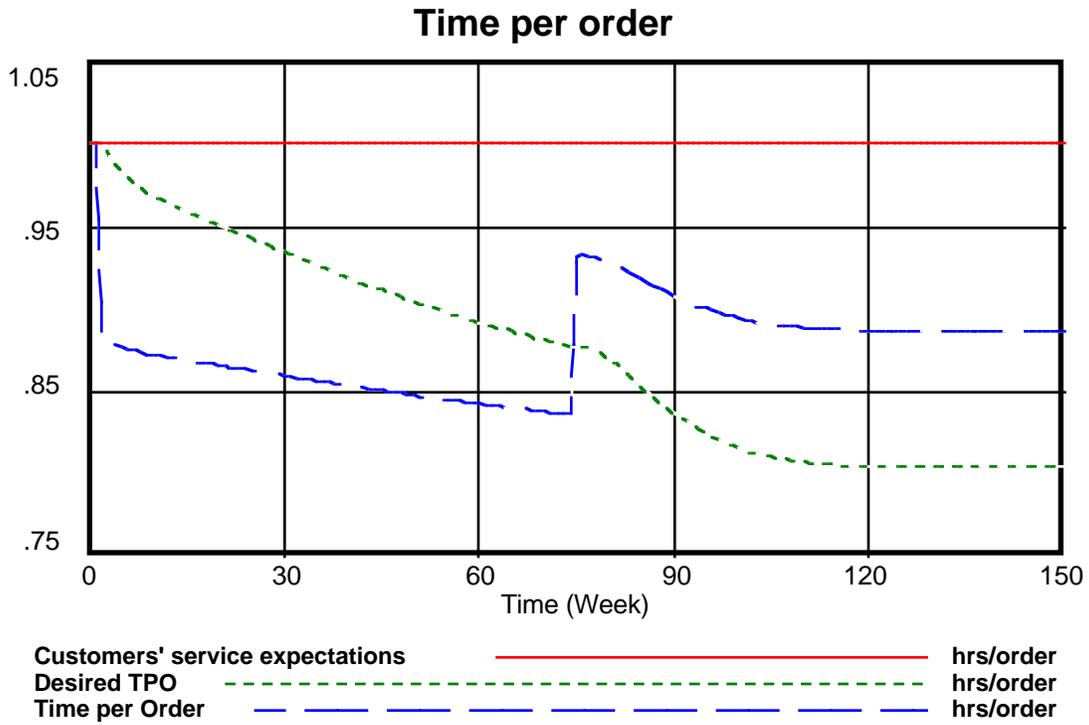
Figure 5.8a   Quality Std. based on past performance. Desired time per order

Desired time per order -- rates

| adjustment of dtpo | ———————————— | (hrs/order)/week |
| erosion of dtpo | - - - - - - - - - - - - - - | (hrs/order)/week |
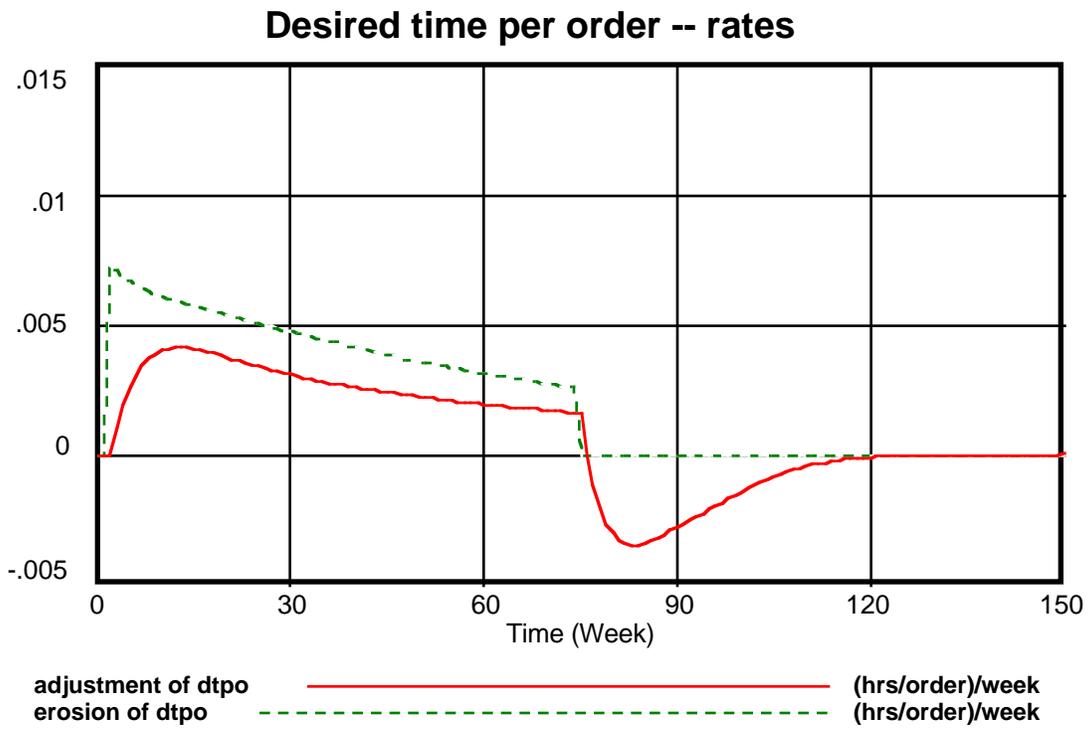
Figure 5.8b   Quality Std. based on past performance. Desired time per order – rates

When work pressure is removed from the system (week 75) the erosion rate returns to zero. However, the lower quality standard forces the adjustment process to drive the desired time-per-order down. The negative adjustment of desired time per order stops when the perception of delivered quality adjusts the employees' quality standard up and the system reaches equilibrium.

The two structures analyzed above (fixed quality standard and quality standard formed by adjusting to past performance) represent the extremes of a continuum of possible structures to use and respond to service quality information. While adjusting the quality standard to past performance triggers erosion of service quality, having a quality standard that is fixed and insensitive to changes in customer requirements might result in poor performance or excessive cost. Where a service setting is positioned in this continuum depends on the use it makes of the additional feedback mechanisms to regulate quality pressure (managerial goal and feedback from market surveys).

Having identified the leverage points for sustaining desired time per order (maintaining quality standard and increasing the effect of quality pressure on time per order), the challenge becomes how to operationalize them in the service delivery context.

`Maintaining the quality standard.` If there is no professional quality standard in the industry, the only option available is for management to have an active role in the formation of the quality standard by explicitly providing an operational quality goal (G). As discussed in chapter three, the managerial quality goal is also vulnerable to erosion based on past performance, but its longer adjustment time constant provides a second line of defense for the erosion of the quality standard.

`Effect of quality pressure on time per order.` Increasing the effect of quality pressure would require management to become aware of the implications of a reduced time per order, i.e., lower sales, and inform employees of those opportunity costs. For example, although loan officers in Nelson House reported some discomfort with their quality performance (evidence of quality pressure) it was not possible to identify any effects from quality pressure on the formation of desired time per order (γ) – even during the periods of low work pressure. The reason for this lack of effect was the misinterpretation that the reduction of time per order had no impact on the LC's performance and that customers would eventually get used to the reduced contact.

Time per order

Customers' service expectations —————————————————— hrs/order
Desired TPO --- hrs/order
Time per Order — —— —— —— —— —— —— —— —— hrs/order
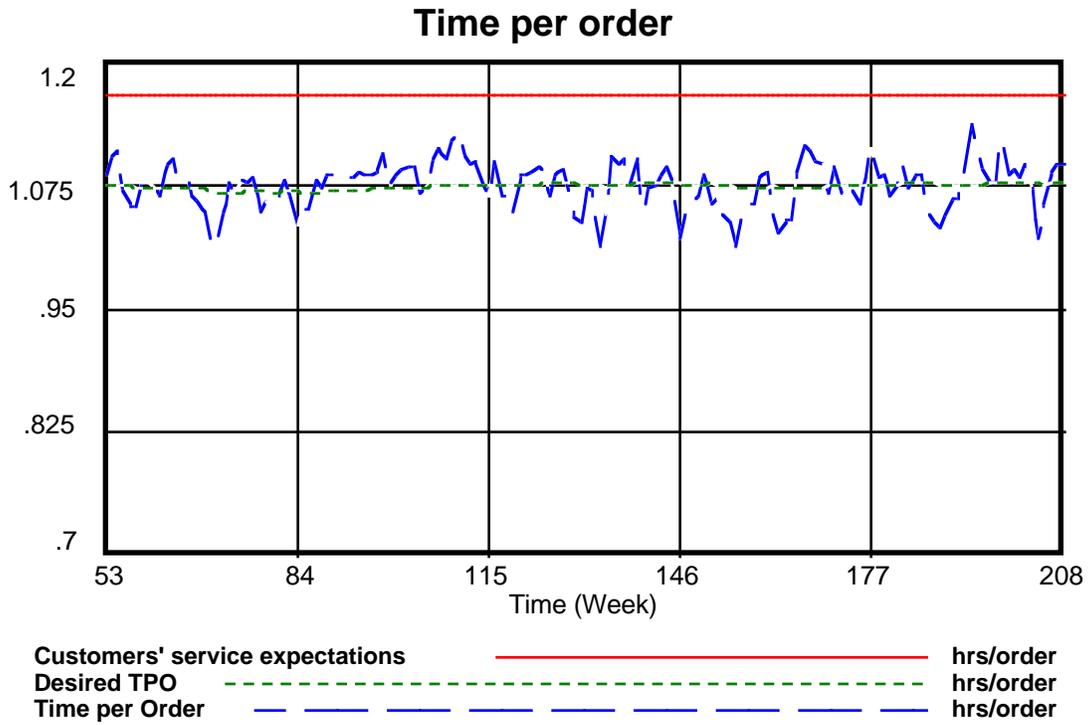
Figure 5.9a   Recommended policies. Desired time per order

Quality

Delivered quality ————————————————————— Quality
Emps' perception of quality ----------------------------- Quality
Quality standard — —— —— —— —— —— —— —— — Quality
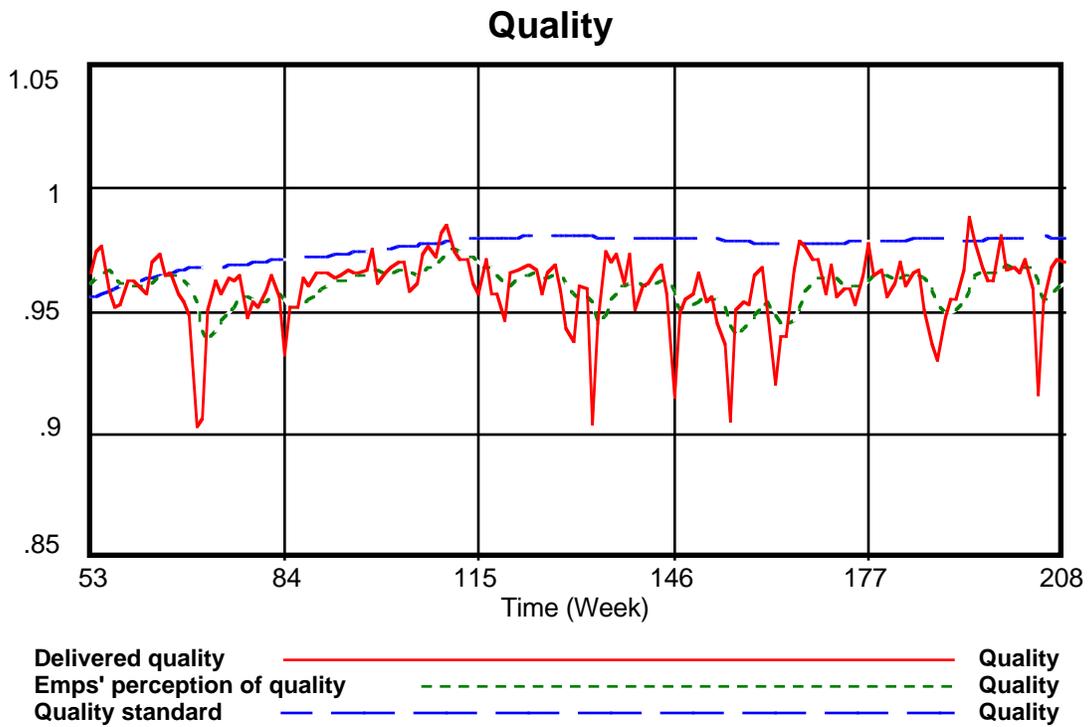
Figure 5.9b   Recommended policies. Service quality

Figures 5.9 show simulated behavior of time per order and service quality in a system with the characteristics of the Nelson House LC and the policy recommendations derived in this section. The effect of work pressure on time per order has been reduced 50% to -0.32, the formation of the quality standard now has a 50% component based on a fixed managerial goal ($\xi = 1.0$), and the effect of quality pressure on time per order has been set to 0.5, with a time constant for the adjustment of desired time per order equal to the time constant of the erosion rate.

The system was initialized in full equilibrium with the exception of the quality standard which was initialized at the current level of service delivery. Although during the first year of the simulation the internal quality standard adjusted towards the managerial quality goal, the overall system proved to be very resistant to policy changes. The policies were capable of sustaining the desired time per order, but the gap towards customer's service expectations (the managerial goal) was not closed in three simulated years.

§5.4. Beyond High-Contact Service Settings

The policy recommendations derived above are specific to the challenge of managing service quality in high-contact services. As we move away from the characteristics of high-contact services (high customization and training requirements) different structural components of the service delivery process may become more important. Changes in dominant structure affect the dynamic behavior of the system, and managerial concerns shift accordingly. To assess the transferability of insights and recommendations derived for the high-contact service sector we need to explicitly address the issue of external validity of the theory – "the extent to which one can generalize the results of a research to the populations and settings of interest" (Judd, Smith and Kidder, 1991, pg. 28).

External validity can be explored in two dimensions: the range of behaviors and reference modes that the theory is capable of explaining, and the variety of service settings that can accurately be captured by the proposed structure. The two dimensions – behavior and structure – define the application domain of the theory.

The variety of reference modes that can be generated by the model has been partially addressed in chapter three. This section identifies the main characteristics of service settings that can be captured in the model and the reference modes that could be expected from the structural characteristics of the service setting. The identification of the main

characteristics of service delivery process not only allows exploration of the flexibility of the model to capture other service settings, but it also permits the identification of the characteristics that define the space where particular policy recommendations are valid.

§5.4.1. Service Delivery Dimensions

Recognizing the enormous variety of processes that can be grouped into the generic label 'service' and the need to generate more specific strategies for managing the service delivery process, the operations management and marketing literatures have gone to great lengths to classify service settings according to dominant characteristics – see Lovelock (1983) for a summary of classifications. In fact, the category high-contact service used throughout this dissertation is just one of such classifications under the dimensions of high customization and high labor intensity (Schmenner, 1986). The proposed service delivery model, with the multiple characteristics of the service delivery process that it captures, can be taken outside the high-contact service context and used to derive insights and policy recommendations for other types of services.

For example, following Lovelock (1983), one can think of the space that defines the service transaction in terms of the degree of judgment exercised by customer-contact-personnel and the extent to which service characteristics are customized for each transaction. These characteristics can be captured operationally by the proposed model. Service settings where judgment plays an important role will normally require personnel with more elaborated training (low initial effectiveness and longer time to become fully effective), and the need for customization requires a flexible service delivery process. While high training requirements reduce the flexibility of the policies to acquire service capacity (because of the time required to train personnel), the need for customization inhibits the standardization of the service delivery process suggesting stronger effects of work pressure on time per order and higher rates of erosion of service quality.

Other dimensions of the service delivery process that can be captured in the model are the time it normally takes for a transaction to be processed – from the rapid exchange that takes place in a fast food restaurant to the elaborated and time consuming negotiations that one could have with a lawyer – and the capital-intensity of the service delivery process. Similarly, one can capture through the model parameters the main characteristics of managerial decisions and the behavioral responses to the various pressures on the delivery process. Table 5.4 shows the service characteristics that are captured by the model and make it customizable to different settings.

Service process characteristics	Personnel characteristics	Managerial characteristics
• Technology • Time constraints • Training requirements • Learning curve • Customization requirements	• Response to work pressure • Response to quality pressure • Effects of fatigue • Turnover • Formation of quality standard	• Formation of work pressure • Effects of work pressure • Formation of quality pressure • Capacity acquisition

Table 5.4   Service delivery characteristics

The next sub-section establishes the link between the structural components of the service delivery process and the behavior that could be expected from such structure.

§5.4.2. Effect of Structure on System Behavior

The structural characteristics identified above have a direct impact on the strength of the different responses to work pressure (see table 5.1). For example, the use of the reduction of time per order as a way to deal with work pressure will be limited if the service delivery process has been standardized, or if professional quality standards constrain the customer-facing personnel to provide a certain service level. The work intensity response is limited by the working hours in the service setting and the willingness and incentives that employees have to work overtime. Finally, the responsiveness of changes in service capacity is limited by the amount of training required, the speed at which additional service capacity could be acquired, and the managerial policies in place. A detailed list of the main constraints on each of the response mechanisms is presented in table 5.5. Again, each of the limiting factors can be related to one of the structural characteristics of the service setting and into specific parameter values within the model.

Having mapped the structural characteristics of the service setting into the basic responses to work pressure, it is possible to identify the relative strength of the responses to work pressure that could be expected from those structural components. For example, in a service with a standardized (rigid) delivery process and relative short training requirements, e.g., a fast-food restaurant, the relative strength of the responses to work pressure that could be expected is: first, increase of work intensity (WI), then increase service capacity (SC), and, probably very weakly, a reduction in the time-per-order (TPO). As a shorthand the ranking will be denoted WI>SC>TPO. It is worth noting that other services, with different structural characteristics, could show the same ranking of responses, e.g., a capital intensive service, such as utility, that is forced to use equipment

beyond the natural maintenance cycle to satisfy an increase of demand. Figure 5.10 presents an example of a service setting that has each of the possible combinations of response flexibility.

Response	Factors limiting flexibility of response
Time per Order	• Standardized service delivery process (low customization) • Professional quality standard (high customization) • Quality sensitive customers • Good information of quality performance
Work intensity	• Constraint on working hours • High customer contact time • Regulated work-hours (airline pilots) • Employee's lack of empathy with customers
Service capacity	• Capital intensity • Long training requirements • Professional certification • Employees union • Long hiring delay • Financial Pressures

Table 5.5   Factors limiting the flexibility of responses to work pressure
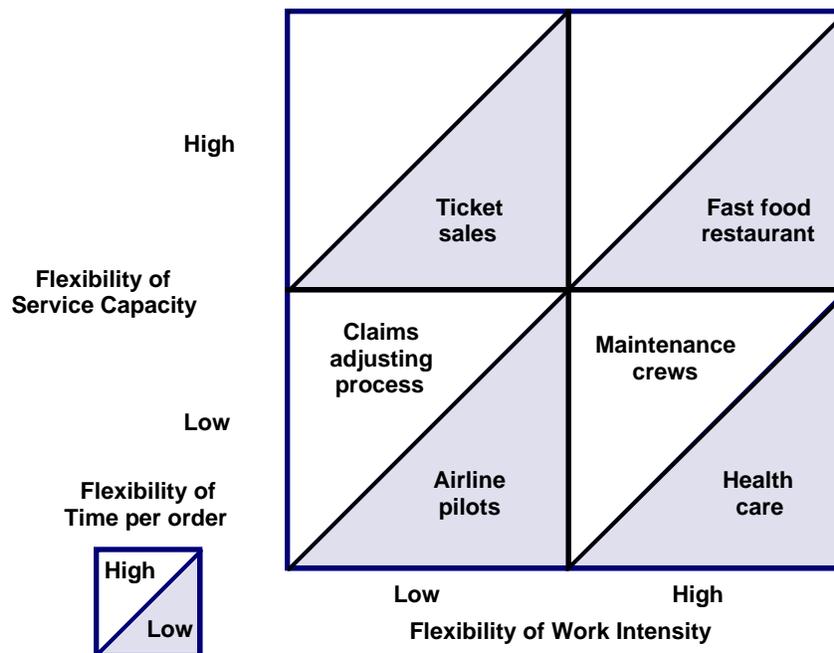
Figure 5.10   Combinations of response flexibility. Examples[9]

Because the similar response rankings create similar dynamic behaviors, regardless of the structural limitations causing the ranking, the relative strength of the responses to work

---

[9] The combination of high flexibility of service capacity with high flexibility of time per order is not feasible since high customization (flexibility in TPO) implies long training requirements.

pressure could be used as a way to classify the reference modes that characterize services. Such classification allows for a reduction of the dimensional characteristics needed to differentiate service settings, and creates a direct linkage between the dominant structural characteristics of the service setting and its dynamic behavior. The insights for management might come from designing policies capable of dealing with the unintended detrimental behaviors that emerge from the different combinations of responses to work pressure.

Table 5.6 lists some examples of service settings with the main structural characteristics (second column) that define the ranking of the responses to work pressure (third column). The fourth column lists the unintended dynamics that could be expected from such response ranking and the last column lists references in the literature that either report observations of the described reference mode or have articulated hypotheses to explain it.

Example	Structural characteristics	Relative response	Expected behavior	References
Airline pilots	• Standardized service delivery process • Regulated work hours • Long training requirements	SC>WI&TPO	Excess capacity	(Sasser, 1976)
Ticket sales	• Standardized service delivery process • Limited work hours • Low training requirements	SC>WI&TPO	Long delivery delays	(Larson, 1987; Maister, 1985)
Health care	• Strict professional quality standards • High empathy with customers • Long training/capital intensive	WI>TPO>SC	Employee burnout and erosion of service quality	(Farber, 1983; Golembiewski, Munzenrider and Carter, 1983; Levin, Roberts et al., 1976)
Fast food restaurant	• Standardized service delivery process • Low face-to-face contact • Low training requirements	WI>SC>TPO	High turnover	(Schlesinger and Heskett, 1991)
Claims adjusting process	• High customization • Limited work hours • Long training requirements	TPO>WI>SC	Erosion of service quality	(Senge and Sterman, 1992)

Table 5.6   Effect of structural components on system behavior[10]

---

[10] The combination TPO>SC>WI was not found feasible. Most employers would allow overtime before incurring in additional service capacity.

§5.4.3. Examples

This section illustrates the model's flexibility by adapting its structural characteristics to two very different service settings: a fast food restaurant – with a highly standardized service delivery process and low training requirements – and health care provision that is highly customizable and has strict professional certification and quality standards.

Figures 5.11a and 5.11b show the response of a system calibrated to the characteristics of a fast food restaurant to a 10% increase in a stationary random series of customer orders[11]. To simplify the comparison of results the calibration to the fast food restaurant was based on the values of the parameters estimated for the LC with three modifications: 1) the average time to achieve full productivity was reduced to 4 weeks, 2) the average time for turnover was reduced to 6 months, and 3) the relative responses to work pressure were modified. Because of the standardized process it was assumed that there was no effect of work pressure in time per order ($\alpha = 0$), and the response of work intensity was modified from $\beta = 0.36$ to $\beta = 0.5$.

The simulated behavior replicates the high turnover predicted by Schlesinger and Heskett's 'cycle of failure in service' theory (1991) for low-skill service settings. Because of the standardized service delivery process the only immediate option to deal with work pressure is work intensity. Average work intensity affects employees' productivity and in turnover, thus reducing the effective service capacity and generating additional work pressure. Although the simulated turnover faction is higher than normal, the relative impact of this dynamic is minimized because of the short training time (effective labor never drops below 95% of its initial experience mix).

The model, however, does not reach a new equilibrium. To generate the new equilibrium proposed by the 'cycle of failure in services' it is necessary to make the turnover rate dependent on technology and the scope of the job design. Since these are assumed constant in the simulation, and the model makes the assumption that the turnover rate is only dependent on service quality and the effects of work intensity the system is brought back to the original equilibrium through increasing service capacity. Management response of increasing service capacity, takes two years to bring the system back to equilibrium.

---

[11] The model was initialized in equilibrium. The stationary series was generated using the pink nois macro (see footnote 34 in Chapter 4) with a normalized standard deviation of 3.2% and a smoothing time constant of 1 week.

Response to work pressure

TPO response	————————	Dimensionless
WI response	– – – – – – – –	Dimensionless
SC response	— — — — — — —	Dimensionless
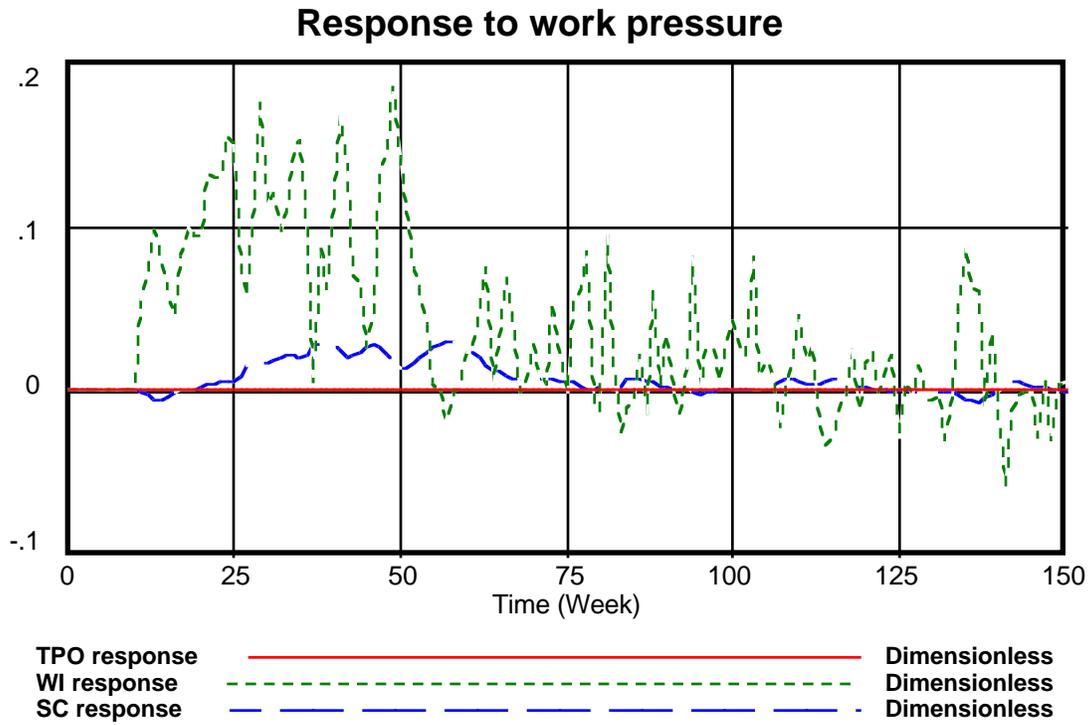
Figure 5.11a   Fast food restaurant. Relative response to work pressure

Long Term Response (Normalized)

| Desired time per order | ———————— | Percent |
Service capacity	– – – – – – – –	Percent
Effective Labor	— — — — — — —	Percent
Throughput	— · — · — · — · —	Percent
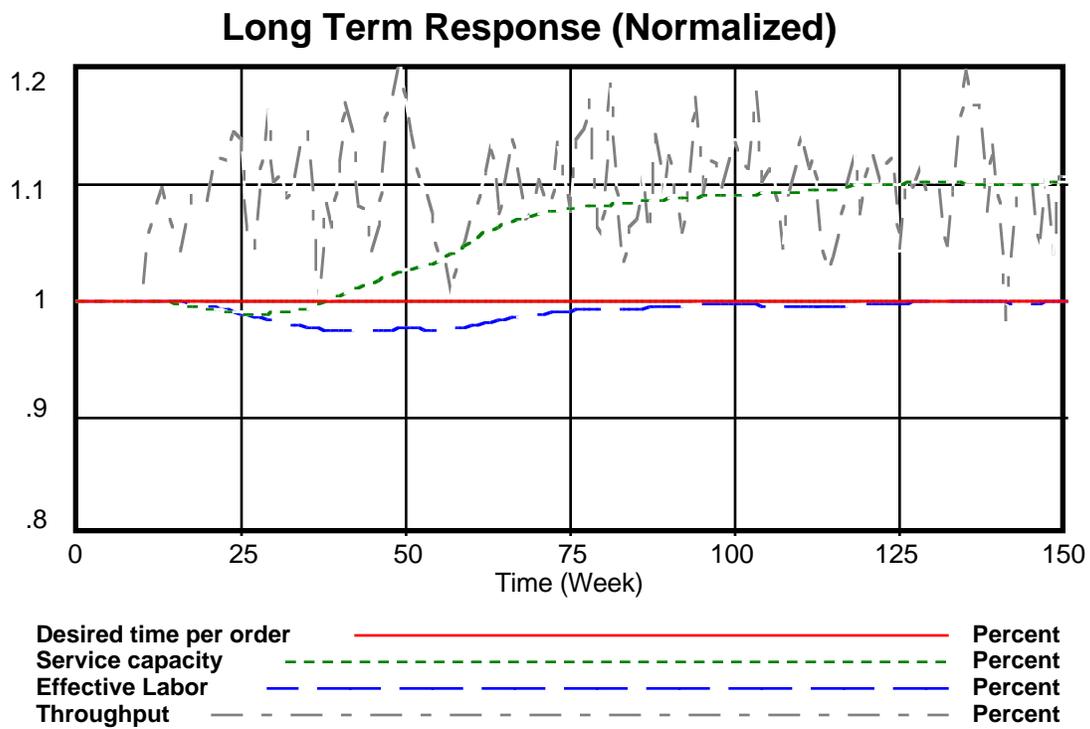
Figure 5.11b   Fast food restaurant. Integrated long-term response

Figures 5.12a and 5.12b show the response of a system calibrated to the characteristics of a health care provided to the same input test described above. The calibration, again, was based on the LC parameters with the following changes: 1) the average time to achieve full productivity was extended to two years, 2) time for turnover was extended to 20 years and the impact of work intensity on turnover was reduced by half, 3) the formation of quality pressure was based on a fixed professional standard and the impact of quality pressure on time per order was increased ($\gamma = 0.5$) through a symmetrical adjustment process of the desired time per order, and 4) the responses to work pressure were modified to give higher priority to work intensity ($\alpha = -0.2$ and $\beta = 0.5$).

Because of the limited adjustment of desired time per order – low adjustment and strict professional quality standards – and the low flexibility to acquire service capacity – long training requirements – most of the response to work pressure falls into work intensity. Sustained work intensity generates the burnout pattern that has been well documented in the human services (Farber, 1983; Golembiewski, Munzenrider and Carter, 1983; Levin, Roberts et al., 1976). Although work intensity represents the most aggressive response to work pressure and there are tight quality standards on the service delivery the delays in bringing in additional service capacity eventually eroded the desired time per order. Although the system had not reached equilibrium in the displayed simulation horizon extended simulations showed that the overshoot in service capacity (caused by the learning curve) is not enough to bring back the desired time per order back to its initial value.

The above discussion bears to the flexibility of the proposed model for service delivery to capture a variety of service settings, and allows the identification of the structural characteristics of settings where erosion of quality standard could be expected – services where TPO>SC. Although the model seems to have the potential to be calibrated to other settings and capture other problematic reference modes in the service industry (desirable characteristics of a generic structure), the insights and recommendations generated in this chapter should only be considered valid within high-contact settings.

Response to work pressure

TPO response	**Dimensionless**
WI response	**Dimensionless**
SC response	**Dimensionless**
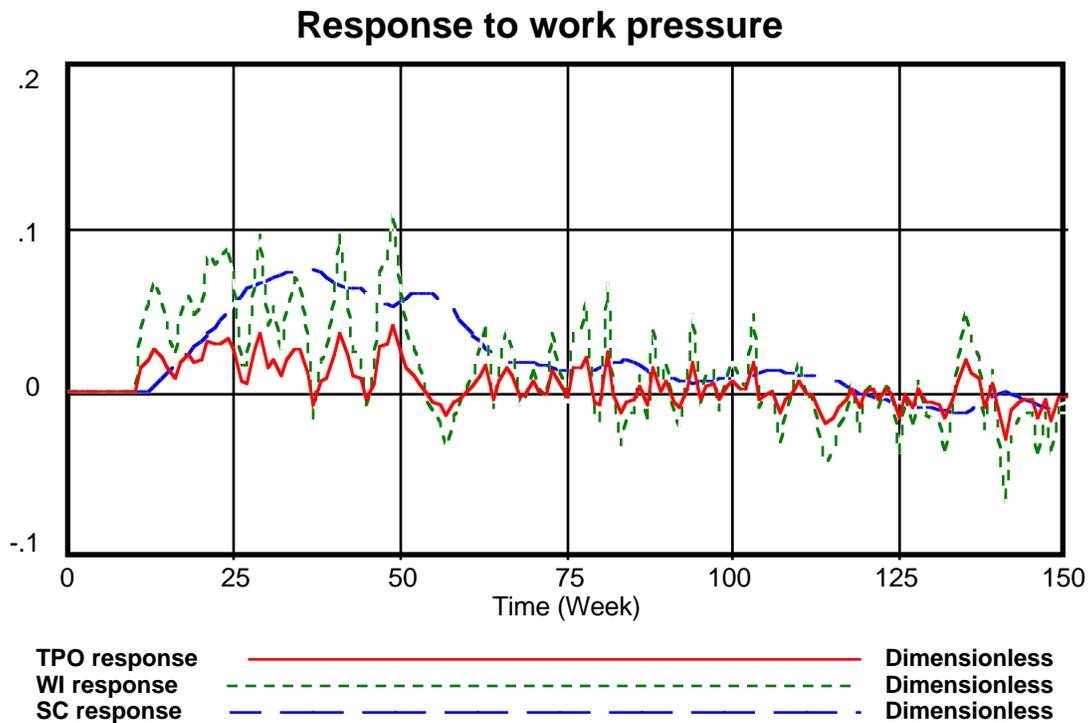
Figure 5.12a   Health care services. Relative response to work pressure

Long Term Response (Normalized)

Desired time per order	**Percent**
Service capacity	**Percent**
Effective Labor	**Percent**
Throughput	**Percent**
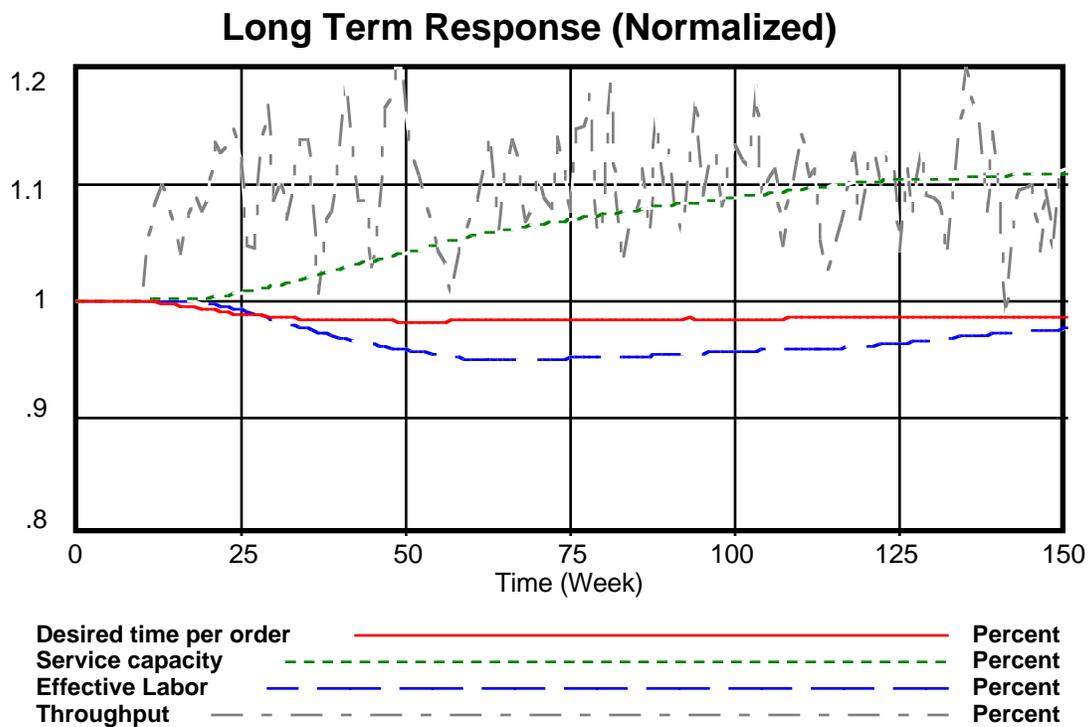
**Figure 5.12b   Health care services. Integrated long-term response**

§5.5. Conclusions

This chapter presented an exploration of the response rate of the various control mechanisms available to deal to high work pressure and presented some evidence about their pervasiveness in the high contact service industry. Summarizing, high-contact services are characterized by high training requirements and the need to customize each service transaction. While high training requirements reduce the flexibility of the policies to acquire service capacity, the need for customization inhibits the standardization of the service delivery process, allowing service employees to reduce the service scope as a response to work pressure. The combination of these characteristics ranks the relative responses that the service setting has to work pressure in a way that biases the service setting towards the erosion of service quality.

Based on these findings, and a revised formulation of the formation of desired time-per-order, a set of policy recommendations to avoid the erosion of service quality were generated and tested. It was shown that a successful strategy for sustaining service quality should reduce work pressure, while simultaneously generating quality pressure and translating it into operational guidelines for allocation of time-per-order.

Finally, the model was taken it beyond high-contact service settings. The structural components determined by the various dimensions of the service delivery process were qualitatively linked to existing theories and empirical descriptions of behavioral modes in other service settings. The exploration of the application domain of the model identified the range of conditions where the policy recommendations are applicable and generated a new framework to link structural characteristics of service setting to expected dominant behaviors in the industry.

§5.6. Future Research Directions

The findings and limitations of this work point to three separated areas where future research should be carried out.

1. Further validation of the theory

- Explore the new proposed formulation for the formation of the desired time-per-order goal in a place where some quality pressure is in place; perhaps in a service setting with low customization.

- Increase confidence and/or update the proposed model structure by replicating the calibration analysis in other service settings, in both high-contact services and settings with other structural characteristics.

2. Extend existing model.

The model boundary needs to be expanded to include other structural characteristics of service settings and causal relations that were excluded from this first iteration. Among other changes, it could be extended to:

- Formally introduce in the demand for production factors the feedback from financial pressures.

- Introduce into the model market dynamics; especially the life value of a customer (reorders) as a function of delivered quality.

- Formalize the link between quality of work life (as function of the technological content of the service delivery process) and turnover.

- Expand the dimensions of service quality beyond time per order.

3. Formalize and generalize empirical findings

- The ranking of the relative strengths of responses to work pressure seems to be a promising framework to understand the challenges of managing service delivery and quality. Further research into the structural characteristics generating such rankings and the unanticipated detrimental behaviors emerging from each ranking could provide insights and recommendations for management beyond the high-contact service context.

- The multiple feedback mechanisms available for management to sustain quality pressure (management goal and feedback from market research) deserve more careful considerations into the creation of policy recommendations.

References

Abdel-Hamid, T.K. and S.E. Madnick. 1991. *Software Project Management Dynamics: An Integrated Approach.* Englewood Cliffs, NJ: Prentice Hall.

Aranda, R., T. Fiddaman and R. Oliva. 1993. Quality Microworlds: Modeling the Impact of Quality Initiatives over the Software Product Life Cycle. *American Programmer*, **6** (5), 52-61.

Arrow, K.J., H.B. Chenery, B. Minhas and R.M. Solow. 1961. Capital-Labor Substitution and Economic Efficiency. *Review of Economics and Statistics*, **43** (5), 225-254.

Austin, J.L. 1962. *How to Do Things with Words.* Cambridge, MA: Harvard University Press.

Band, W.A. 1991. *Creating Value for Customers.* New York: John Wiley & Sons, Inc.

Barlas, Y. and S. Carpenter. 1990. Philosophical roots of model validation: two paradigms. *System Dynamics Review*, **6** (2), 148-166.

Baumol, W.J. 1967. Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis. *The American Economic Review*, **57** (June), 415-426.

Baumol, W.J., S.A.B. Blackman and E.N. Wolff. 1985. Unbalanced Growth Revisited: Asymptotic Stagnancy and New Evidence. *American Economic Review*, **75** (4), 806-817.

Bell, D. 1973. *The Coming of Post-Industrial Society: A Venture In Social forecasting.* New York: The Basic Books.

Bell, J.A. and J.F. Bell. 1980. System Dynamics and Scientific Method. In J. Randers (Ed.), *Elements of the System Dynamics Method.* (pp. 3-22). Cambridge, MA: Productivity Press.

Bell, J.A. and P.M. Senge. 1980. Methods for Enhancing Refutability in System Dynamics Modeling. *TIMS Studies in the Management Sciences*, **14** (1), 61-73.

Bitran, G.R. and M.P. Lojo. 1993a. A Framework for Analyzing Service Operations. *European Management Journal*, **11** (3), 271-282.

Bitran, G.R. and M.P. Lojo. 1993b. Framework for Analyzing the Quality of Customer Interface. *European Management Journal*, **11** (4), 385-396.

Boulding, W., R. Staelin, A. Karla and V.A. Zeithaml. 1992. *Conceptualizing and Testing a Dynamic Process Model of Service Quality*. Marketing Science Institute. Cambridge, MA. August 1992.

Britting, K.R. 1973. *Correlated Noise Generation Using DYNAMO*. System Dynamics Group, MIT. D-1908.

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

Burchill, G.W. 1993. *Concept Engineering: An investigation of TIME vs. MARKET orientation in product concept development*. PhD Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Buzzell, R.D. and B.T. Gale. 1987. *The PIMS Principles: Linking Strategy to Performance.* New York: The Free Press.

Campbell, D.T. and D.W. Fiske. 1959. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, **56** (2), 81-105.

Cannell, C.F. and R.L. Kahn. 1968. Interviewing. In G. Lindzey and E. Aronson (Ed.), *The Handbook of Social Psychology.* (pp. 526-595). Reading, MA: Addison-Wesley Publishing Co.

Cavaleri, S. and J.D. Sterman. 1995. Towards Evaluation of Systems Thinking Interventions: A Case Study. In T. Shimada and K. Saeed (Ed.), *1995 International System Dynamics Conference*, (pp. 398-407). Tokyo, Japan.

Central Statistical Office. 1995. *Regional Trends: 30th year of Regional Statistics.* London: Government Statistical Service.

Chase, R.B. 1981. The Customer Contact Approach to Services: Theoretical Bases and Practical Extensions. *Operations Research*, **29** (4), 698-706.

Chase, R.B. and N.J. Aquilano. 1989. *Production and Operations Management: A Life Cycle Approach.* (5th Ed.). Homewood, IL: Irwin.

Chase, R.B. and J.L. Heskett. 1995. Introduction to the Focused Issue on Service Management. *Management Science*, **41** (11), 1717-1719.

Chase, R.B. and D.A. Tansik. 1983. The Customer Contact Model for Firm Design. *Management Science*, **29** (9), 1037-1050.

Checkland, P.B. 1985. From Optimizing to Learning: A Development of Systems Thinking for the 1990s. *Journal of the Operational Research Society*, **36** (9), 757-767.

Churchman, C.W., R.L. Ackoff and E.L. Arnoff. 1957. *Introduction to Operations Research.* New York: John Wiley & Sons, Inc.

Cicourel, A. 1974. *Cognitive Sociology: Language and Meaning in Social Interaction.* New York: The Free Press.

Clark, C. 1940. *The Conditions of Economic Progress.* London: Macmillan.

Cohen, S.S. and J. Zysman. 1987. *Manufacturing Matters: The Myth of a Post-Industrial Economy.* New York: Basic Books, Inc.

Cook, T.D. and D.T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston: Houghton Mifflin Company.

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

Delaunay, J.C. and J. Gadrey. 1992. *Services in Economic Thought: Three Centuries of Debate.* Boston: Kluwer Academic Publishers.

Deming, W.E. 1982. *Out of the Crisis.* Cambridge, MA: MIT Press.

Denzin, N.K. 1970. Triangulation: A Case for Methodological Evaluation and Combination. Part XII. In N.K. Denzin (Ed.), *Sociological Methods: A Sourcebook.* (pp. 469-522). Chicago: Aldine Publishing.

Déry, R., M. Landry and C. Banville. 1993. Revisiting the issue of model validation in OR: An epistemological view. *European Journal of Operational Research*, **66** (2), 168-183.

Diewert, W.E. 1971. An Application of the Shepard Duality Theorem: A Generalized Linear Production Function. *Journal of Political Economy*, **79** (3), 482-507.

Drewes, W.F. 1991. *Quality Dynamics for the Service Industry.* Milwaukee, WI: ASQC Quality Press.

Ehrengerg, R.G. 1971. *Fringe Benefits and Overtime Behavior.* Lexington, MA: Lexington Books.

Einhorn, H.J. and R.M. Hogarth. 1981. Behavioral Decision Theory: Process of Judgment and Choice. *Annual Review of Psychology*, **32**, 53-88.

Farber, B.A. (Ed.). 1983. *Stress and Burnout in the Human Service Professions.* New York: Pergamon Press.

Feinberg, M.R. and A. Levenstein. 1985. "It's Not my Job, Man". *Wall Street Journal*, November 11, 1985.

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

Forrester, J.W. 1968. A Response to Ansoff and Slevin. *Management Science*, **14** (9), 601-618.

Forrester, J.W. 1969. *Urban Dynamics.* Cambridge, MA: MIT Press.

Forrester, J.W. 1973. *Confidence in Models of Social Behavior, With Emphasis on System Dynamics Models*. System Dynamics Group, MIT. Memo D-1967.

Forrester, J.W. 1975. The Impact of Feedback Control Concepts on the Management Sciences. In *Collected Papers of Jay W. Forrester.* (pp. 45-60). Cambridge, MA: Productivity Press.

Forrester, J.W. 1979. An Alternative Approach to Economic Policy: Macrobehavior from Microstructure. In N.M. Kamrany and R.H. Day (Ed.), *Economic Issues of the Eighties.* (pp. 80-108). Baltimore, MD: The John Hopkins University Press.

Forrester, J.W. 1993. System Dynamics and the Lessons of 35 Years. In K.B. De Greene (Ed.), *Systems-Bases Approach to Policymaking.* (pp. 199-240). Norwell, MA: Kluwer Academic Publishers.

Forrester, J.W. and N.J. Mass. 1976. *The production Sector of the System Dynamics National Model: Equation Description*. System Dynamic Group, Massachusetts Institute of Technology. Cambridge, MA. Working paper D-2486.

Fuchs, V.R. 1968. *The Service Economy.* New York: National Bureau of Economic Research.

Gass, S.I. 1983. Decision-Aiding Models: Validation, Assessment and Related Issues for Policy Analysis. *Operations Research*, **31** (4), 603-631.

Gershuny, J.I. 1978. *After Industrial Society? The Emerging Self-Service Economy.* Atlantic Highlands, NJ: Humanities Press.

Golembiewski, R.T., R. Munzenrider and D. Carter. 1983. Phases of Progressive Burnout and Their Work Site Covariants: Critical Issues in OD Research and Praxis. *Journal of Applied Behavioral Science*, **19** (4), 461-481.

Goodwin, R. 1951. The Nonlinear Accelerator and the Persistence of Business Cycles. *Econometrica*, **19** (1), 1-17.

Graham, A.K. 1980. Parameter Estimation in System Dynamics Modeling. In J. Randers (Ed.), *Elements of the System Dynamic Method.* (pp. 143-161). Cambridge, MA: Productivity Press.

Gronroos, C. 1984. A Service Quality Model and its Marketing Implications. *European Journal of Marketing*, **18** (4), 36-44.

Gummesson, E. 1993. *Quality Management in Service Organizations.* Stockholm: International Service Quality Association.

Harker, P.T. 1995. Introduction: Service-Sector Productivity—The MS/OR Challenge. *Interfaces*, **25** (3), 1-5.

Haywood-Farmer, J. 1988. A Conceptual Model of Service Quality. *International Journal of Operations and Production Management*, **8** (6), 19-29.

Herzberg, F. 1966. *Work and the Nature of Man.* New York: Thomas Y. Crowell.

Heskett, J.L., T.O. Jones, G.W. Loveman, W.E. Sasser and L.A. Schlesinger. 1994. Putting the Service-Profit Chain to Work. *Harvard Business Review*, **72** (2), 164-174.

Hines, J.R. 1987. *Essays in Behavioral Economic Modeling.* PhD Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Hoech, J. 1988. *Quality Frameworks in the Service Industry*. MS Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Hogarth, R.M. 1980. *Judgment and Choice: The Psychology of Decision.* New York: John Wiley & Sons, Inc.

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

Hostage, G.M. 1975. Quality Control in a Service Business. *Harvard Business Review*, **53** (4), 98-106.

Huber, P. 1987. Injury Litigation and Liability Insurance Dynamics. *Science*, **238**, 31-36.

Ingle, S. and N. Ingle. 1983. *Quality Circles in Service Industries.* Englewood Cliffs, NJ: Prentice Hall.

Internal Revenue Service. 1967. *Statistics of Income, Corporation Income Tax Returns.* Washington, DC: Department of the Treasury.

Jackson, S.E., R.L. Schwab and R.L. Schuler. 1986. Toward an Understanding of the Burnout Phenomenon. *Journal of Applied Psychology*, **71** (4), 630-640.

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

Joseph, W. 1983. *Professional Service Management.* New York: McGraw-Hill Book Company.

Judd, C.M., E.R. Smith and L.H. Kidder. 1991. *Research Methods in Social Relations.* Fort Worth, TX: Holt, Rinehart and Winston, Inc.

Kahneman, D. and A. Tversky. 1982. The Psychology of Preferences. *Scientific American*, **246**, 160-173.

Kano, N., N. Seraku, F. Takahashi and S. Tsuji. 1984. Attractive Quality and Must-be Quality (1), (2). *Journal of Japanese Society of Quality Control*, **14** (2), 1-12.

Kim, D.H. 1989. Learning Laboratories: Designing a Reflective Learning Environment. In P. Milling and E. Zahn (Ed.), *1989 International System Dynamics Conference*, (pp. 327-334). Berlin.

Kirschenbaum, A. and J. Weisberg. 1990. Predicting Worker Turnover: An Assessment of Intent and Actual Separations. *Human Relations*, **43** (9), 829-847.

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

Kossoris, M. 1947. *Studies of the Effects of the Long Working Hours*. Bureau of Labor Statistics. Washington, D.C. Bulletins 791 and 791A.

Kuhn, T.S. 1970. *The Structure of Scientific Revolutions.* (2nd Ed.). Chicago: University of Chicago Press.

Lakatos, I. 1974. Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos and A. Musgrave (Ed.), *Criticism and the Growth of Knowledge.* (pp. 91-196). Cambridge: Cambridge University Press.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things.* Chicago: University of Chicago Press.

Landry, M., J. Malouin and M. Oral. 1983. Model Validation in Operations Research. *European Journal of Operational Research*, **59** (1), 64-84.

Lane, D.C. and R. Oliva. 1994. The Greater Whole: Towards a Synthesis of System Dynamics and Soft Systems Methodology. In E. Wolstenholme and C. Monaghan (Ed.), *1994 International System Dynamics Conference*, Problem-solving methodologies (pp. 134-146). Stirling, Scotland.

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

Larson, R.C. 1987. Perspectives on Queues: Social Justice and the Psychology of Queuing. *Operations Research*, **35** (6), 895-905.

Lash, L.M. 1989. *The Complete Guide to Customer Service.* New York: John Wiley & Sons, Inc.

Levin, G., E.B. Roberts, G.B. Hirsch, D.S. Kligler, et al. 1976. *The Dynamics of Human Service Delivery.* Cambridge, MA: Ballinger.

Levine, R.L., M. Van Sell and B. Rubin. 1985. A Model of Burnout in the Work Place. In M.E. Warkentin (Ed.), *1985 International Conference of the Systems Dynamics Society*, (pp. 487-501). Keystone, Colorado.

Levinthal, D. and J.G. March. 1981. A Model of Adaptive Organizational Search. *Journal of Economic Behavior and Organization*, **2** (4), 307-333.

Levitt, T. 1972. Production-line approach to service. *Harvard Business Review*, **50** (5), 41-52.

Levitt, T. 1976. The industrialization of service. *Harvard Business Review*, **54** (5), 63-74.

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

Lovelock, C.H. 1983. Classifying Services to Gain Strategic Marketing Insights. *Journal of Marketing*, **47** (3), 9-20.

Lovelock, C.H. (Ed.). 1992. *Managing Services: Marketing, Operations and Human Resources.* (2nd Ed.). Englewood Cliffs, NJ: Prentice Hall.

Lyneis, J.M. 1980. *Corporate Planning and Policy Design: A System Dynamics Approach.* Portland, OR: Productivity Press.

Main, J. 1981. Toward Service without a Snarl. *Fortune*, March 23, 1981. (pp. 58).

Maister, D.H. 1984. Quality Work Doesn't Mean Quality Service. *The American Lawyer*, **1984** (April), 6-8.

Maister, D.H. 1985. The Psychology of Waiting Lines. In J.A. Czepiel, M.R. Solomon and C.F. Surprenant (Ed.), *The Service Encounter: Managing Employee/Customer Interaction in Service Businesses.* (pp. 113-123). Lexington, MA: Lexington Books.

Mass, N.J. and P.M. Senge. 1980. Alternative Test for Selecting Model Variables. In J. Randers (Ed.), *Elements of the System Dynamic Method.* (pp. 203-223). Cambridge, MA: Productivity Press.

MicroWorlds. 1994. *Service Quality Microworld.* Cambridge, MA: MicroWorlds, Inc.

Mills, P.K. 1986. *Managing Service Industries.* Cambridge, MA: Ballinger Publishing Company.

Miser, H.J. 1993. A foundational concept of science appropriate for validation in operational research. *European Journal of Operational Research*, **66** (2), 204-234.

Mitroff, I. 1972. The myth of objectivity or why science needs a new psychology of science? *Management Science*, **18** (10), B613-B618.

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

Moissis, A.A. 1989. *Decision Making in the Insurance Industry: A Dynamic Simulation Model and Experimental Results*. MS Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Morecroft, J.D.W. 1983. System Dynamics: Portraying Bounded Rationality. *Omega*, **11** (2), 131-142.

Morecroft, J.D.W. 1985. Rationality in the Analysis of Behavioral Simulation Models. *Management Science*, **31** (7), 900-916.

NatWest Bank UKBB. 1992. *West End Region Delivery Strategy Plan 1992-1997*. National Westminster Bank. London. May 1992.

Naylor, T.H. and J.M. Finger. 1967. Verification of Computer Simulation Models. *Management Science*, **14** (2), B92-B101.

Ogata, K. 1990. *Modern Control Engineering.* Englewood Cliffs, NJ: Prentice Hall.

Oliva, R. 1992. *Service Quality Management Flight Simulator: Facilitator's Training Guide*. Organizational Learning Center, Massachusetts Institute of Technology. Cambridge, MA. September, 1992.

Oliva, R. 1993a. *Service Quality Management Flight Simulator: User's Guide*. Organizational Learning Center, Massachusetts Institute of Technology. Cambridge, MA. April, 1993.

Oliva, R. 1993b. *Service Quality-Service Capacity Interactions: Framework for a Dynamic Theory*. Systems Dynamics Group, Massachusetts Institute of Technology. Cambridge, MA. November, 1993. Department memorandum D-4371-2.

Oliva, R. 1995. *A Vensim Module to Calculate Summary Statistics for Historical Fit*. System Dynamics Group, MIT. Memo D-4584.

Oreskes, N., K. Shrader-Frechette and K. Belitz. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, **263**, 641-646.

Parasuraman, A., V.A. Zeithaml and L.L. Berry. 1985. A Conceptual Model of Service Quality and its Implications for Further Research. *Journal of Marketing*, **49** (4), 41-50.

Parasuraman, A., V.A. Zeithaml and L.L. Berry. 1994. *Moving Forward in Service Quality Research: Measuring Different Customer-Expectation Levels, Comparing Alternative Scales, and Examining the Performance-Behavioral Intentions Link*. Marketing Science Institute. Cambridge, MA. 1994. 94-114.

Parasuraman, S. 1982. Predicting Turnover Intentions and Turnover Behavior: A Multivariate Analysis. *Journal of Vocational Behavior*, **21** (1), 111-121.

Peterson, D.W. 1980. Statistical Tools for System Dynamics. In J. Randers (Ed.), *Elements of the System Dynamic Method.* (pp. 143-161). Cambridge, MA: Productivity Press.

Pines, A. and D. Kafry. 1978. Occupational Tedium in the Social Services. *Social Work*, **23** (6), 499-507.

Popper, K. 1959. *The Logic of Scientific Discovery.* New York: Basic Books, Inc.

Powell, M.J.D. 1969. A method for non-linear constraints in minimization problems. In R. Fletcher (Ed.), *Optimization.* (pp. 283-293). New York: Academic Press.

Powell, M.J.D. 1972. Problems Related to Unconstrained Optimization. In W. Murray (Ed.), *Numerical Methods for Unconstrained Optimization.* (pp. 29-58). New York: Academic Press.

Quinn, J.B. and C.E. Gagnon. 1986. Will Services Follow Manufacturing into Decline? *Harvard Business Review*, **64** (6), 95-103.

Richardson, G.P. and A.L. Pugh. 1981. *Introduction to System Dynamics Modeling with Dynamo.* Cambridge, MA: MIT Press.

Rosander, A.C. 1989. *The Quest for Quality in Services.* Milwaukee, WI: Quality Press.

Roy, B. 1993. Decision science or decision-aid science? *European Journal of Operational Research*, **66** (2), 184-203.

Sasser, E.W., R.P. Olsen and D.D. Wyckoff. 1978. *Management of Service Operations.* Boston: Allyn and Bacon.

Sasser, W.E. 1976. Match Supply and Demand in Service Industries. *Harvard Business Review*, **54** (6), 133-140.

Sastry, M.A. 1995. *Time and Tide in Organizations: Simulating change processes in adaptive, punctuated, and ecological theories of organizational evolution*. PhD Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Schlesinger, L.A. and J.L. Heskett. 1991. Breaking the Cycle of Failure in Services. *Sloan Management Review*, **32** (3), 17-28.

Schlesinger, L.A. and J. Zornitsky. 1991. Job Satisfaction, Service Capability, and Customer Satisfaction: An Examination of Linkages and Management Implications. *Human Resource Planning*, **14** (2), 141-150.

Schmenner, R.W. 1986. How Can Service Business Survive and Prosper? *Sloan Management Review*, **27** (3), 21-32.

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

Schneider, B. and D.E. Bowen. 1985. Employee and Customer Perceptions of Service in Banks: Replication and Extension. *Journal of Applied Psychology*, **70**, 423-433.

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

Searle, J.R. 1969. *Speech Acts: An Essay on the Philosophy of Language.* London: Cambridge University Press.

Senge, P.M. 1977. Statistical estimation of feedback models. *Simulation*, **28** (June), 177-184.

Senge, P.M. 1978. *The System Dynamics National Model Investment Function: A Comparison to the Neoclassical Investment Function*. PhD Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Senge, P.M. 1985. System Dynamics, Mental Models, and the Development of Management Intuition. In M. Warkentin (Ed.), *1985 International System Dynamics Conference*, (pp. 788-798). Keystone, CO.

Senge, P.M. 1990a. Catalyzing Systems Thinking within Organizations. In F. Masaryk (Ed.), *Advances in Organizational Development.* (pp. 197-246). Norwood, NJ: Ablex.

Senge, P.M. 1990b. *The Fifth Discipline: The Art and Practice of the Learning Organization.* New York: Doubleday Currency.

Senge, P.M. and C. Lannon. 1990. Managerial Microworlds. *Technology Review*, **1990** (July), 63-68.

Senge, P.M. and R. Oliva. 1993. Developing a Theory of Service Quality/Service Capacity Interaction. In E. Zepeda and J.A.D. Machuca (Ed.), *1993 International System Dynamics Conference*, (pp. 476-485). Cancún, México.

Senge, P.M. and J.D. Sterman. 1992. Systems Thinking and Organizational Learning: Acting Locally and Thinking Globally in the Organization of the Future. *European Journal of Operational Research*, **59** (1), 137-150.

Shelp, R.K. 1988. The Service Economy Gets No Respect. In C.H. Lovelock (Ed.), *Managing Services: Marketing, Operations and Human Resources.* (pp. 1-5). Englewood Cliffs, NJ: Prentice Hall.

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

Smith, J.H. 1993. Modeling muddles: Validation beyond the numbers. *European Journal of Operational Research*, **66** (2), 235-249.

Stanback Jr., T.M. 1979. *Understanding the Service Economy: Employment, Productivity, Location.* Baltimore, MD: The John Hopkins University Press.

Sterman, J.D. 1981. *The Energy Transition and the Economy: A System Dynamics Approach*. PhD Thesis, Sloan School of Management, Massachusetts Institute of Technology.

Sterman, J.D. 1984. Appropriate Summary Statistics for Evaluating the Historical Fir of System Dynamics Models. *Dynamica*, **10** (Winter), 51-66.

Sterman, J.D. 1985a. A Behavioral Model of the Economic Long Wave. *Journal of Economic Behavior and Organization*, **6** (2), 17-53.

Sterman, J.D. 1985b. The Growth of Knowledge: Testing a Theory of Scientific Revolutions with a Formal Model. *Technological Forecasting and Social Change*, **28** (2), 93-122.

Sterman, J.D. 1989a. Misperceptions of Feedback in Dynamic Decision Making. *Organizational Behavior and Human Decision Processes*, **43** (3), 301-335.

Sterman, J.D. 1989b. Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment. *Management Science*, **35** (3), 321-339.

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

Summers, R. 1985. Services in the International Economy. In R.P. Inman (Ed.), *Managing the Service Economy. Prospects and problems.* (pp. 27-48). Cambridge: Cambridge University Press.

Theil, H. 1966. *Applied Economic Forecasting.* Amsterdam: North-Holland Publishing Co.

Thomas, H.R. 1993. *Effects of Scheduled Overtime on Labor Productivity: A Literature Review and Analysis*. Pennsylvania State University. University Park, PA. November 1990. Source Document 60.

Thomas, R.J. Unpublished ms. *Doing an Interview*. Center for Research in Social Organization. University of Michigan. Ann Arbor, MI.

Tornow, W.W. 1991. Service Quality and Organizational Effectiveness: Comments from the Guest Editor. *Human Resource Planning*, **14** (2), 86-88.

Tornow, W.W. and J.W. Wiley. 1991. Service Quality and Management Practices: A Look at Employee Attitudes, Customer Satisfaction and Bottom-Line Consequences. *Human Resource Planning*, **14** (2), 105-116.

Tuchman, B. 1980. The Decline of Quality. *New York Times Sunday Magazine*, November 2, 1980. (pp. 38).

Tversky, A. and D. Kahneman. 1974. Judgment Under Uncertainty: Heuristics and Biases. *Science*, **185**, 1124-1131.

Ulrich, D., R. Halbrook, D. Meder, M. Stuchlik and S. Thorpe. 1991. Employee and Customer Attachment: Synergies for Competitive Advantage. *Human Resource Planning*, **14** (2), 89-104.

US Department of Commerce. 1992. *National Income and Product Accounts of the United States. 2 Volumes.* Washington, DC: US Government Printing Office.

US Department of Labor. 1979. *Time Series Data for Input-Output Industries: Output, Price and Employment.* Washington, DC: US Department of Labor. Bulletin 2018.

van Horn, R.L. 1971. Validation of Simulation Results. *Management Science*, **17** (5), 247-258.

Varian, H.R. 1992. *Microeconomic Analysis.* (3rd Ed.). New York: W.W. Norton & Company, Inc.

Vendermerwe, S. 1993. *From Tin Soldiers to Russian Dolls: Creating added value through services.* Oxford: Butterworth-Heinemann, Ltd.

Ventana Systems. 1995. *Vensim 1.62 Reference Manual.* Belmont, MA: Ventana Systems, Inc.

Webb, E.J., D.T. Campbell, R.D. Schwartz and L. Sechrest. 1966. *Unobtrusive Measures: Non Reactive Research in the Social Sciences.* Skokie, IL: Rand McNally.

Wei, W.W. 1990. *Time Series Analysis.* Reading, MA: Addison-Wesley Publishing Co.

Weisberg, J. 1994. Measuring Worker's Burnout and Intention to Leave. *Quality of Working Life*, **15** (1), 4-14.

Wiley, J.W. 1991. Customer Satisfaction and Employee Opinions: A Supportive Work Environment and Its Financial Cost. *Human Resource Planning*, **14** (2), 117-128.

Wyckoff, D.D. 1992. New Tools for Achieving Service Quality. In C.H. Lovelock (Ed.), *Managing Services: Marketing, Operations and Human Resources.* (2nd Ed.). (pp. 236-249). Englewood Cliffs, NJ: Prentice Hall.

Zeithaml, V.A., L.L. Berry and A. Parasuraman. 1987. *Communication and Control Processes in the Delivery of Service Quality*. Marketing Science Institute. Cambridge, MA. June 1987.

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

Zelditch, M.J. 1970. Some Methodological Problems of Field Studies. In N.K. Denzin (Ed.), *Sociological Methods: A Sourcebook.* (pp. 495-522). Chicago: Aldine Publishing.