DEVELOPING A THEORY OF
SERVICE QUALITY/SERVICE CAPACITY INTERACTION

Peter M. Senge and Rogelio Oliva
Center for Organizational Learning, MIT
Room E40-294, 77 Massachusetts Ave.
Cambridge, MA 02139, U.S.A.

ABSTRACT
Service quality cannot be measured and tested in as straightforward a manner as in manufacturing. This bias service businesses to focusing on keeping measurable variables—typically, expenses and work flows—in control, while underinvesting in the intangibles of service capacity and service quality. In the long-term, results can be mediocre levels of service quality, poor customer satisfaction, high turnover of service personnel, and, ultimately, higher total costs. In this paper we will present an emerging theory of interactions between Service Quality and Service Capacity, relate this theory to past research in both the System Dynamics and Total Quality Management traditions, and outline ongoing empirical testing of the theory.

FOCUS OF THE RESEARCH
At the Center for Organizational Learning, we are engaged in developing new tools and methods to help organizations develop systems thinking and related learning capabilities. Present research is focused in three broad substantive areas: managing new product development efforts, understanding determinants of cycle time in complex supply chains, and managing quality in service businesses and developing and implementing "value-adding services" strategies.

In each of these areas, better theories are needed to guide development of better management tools and methods. Each of the areas represents a complex dynamic context in which key variables are interrelated through multiple feedback interactions. Delays, nonlinearities, and non-obvious dynamics confound the efforts of decision makers to bring about improvement through typical management interventions. Without well-developed theories that reveal the dynamic structures underlying problems, managers are prone to focus on problem symptoms rather than areas of high leverage. Without explicit theories, different managers will be guided by different mental models in their improvement efforts. Without explicit theories that can be continuously tested and improved, it is unlikely that interventions will produce new understandings that can be captured and serve as a starting point for further learning.

Current research at the Center for Organizational Learning is focused on developing better theories in each of the three subject areas mentioned above. Specifically, we are developing "system dynamics" computer simulation models and interactive "management flight simulators" embodying generic theories of the dynamics of new product development, complex supply chains, and service quality. These generic theories are being tested in multiple organizational settings as part of the collaborative efforts of the Center with its organization partners. Eventually, these simulators and underlying dynamic models will be disseminated widely to aid continuing research and education.

The purpose of this paper is to lay out an emerging theory of interactions between service quality and service capacity, relate this theory to past research, and outline the empirical tests we plan to conduct of the theory. A subsequent paper will lay out a related theory of implementing value-added services strategies, which is guiding other aspects of the service quality projects.

HISTORICAL DEVELOPMENT OF THE THEORY
The theory of service quality-service capacity interactions that is guiding present research has been developed over several years. The first articulation of the theory emerged in the context of a multiple-
year study with a leading property and liability insurance company. That study\(^1\), which resulted in a "claims management learning laboratory" that operated for several years within the firm, focused attention on the rising costs of claims settlements and litigation and related costs, and the declining overall financial health of the industry. These costs trends have been developing for many years. 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 (such as toxic waste). Our study illuminated internal sources of the problems.

During the last 50 years there has been a rising trend in "loss ratios" (settlement costs and litigation costs relative to premiums) and a falling trend in "expense ratios" for the whole insurance industry (Moissis, 1989). One might interpret the falling expense ratios as evidence of increasing productivity and management innovation. Our work suggests something different. While not denying that there have been some increases in productivity gained from technological innovation in underwriting, claims processing, and customer service, the central hypothesis that has emerged from our work is that the rising settlement costs and the 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 consequence long-term increase in total costs and erosion in profitability have led to increasing focus on expense control and "productivity" (normally defined as customers served per service person), thereby reinforcing underinvestment in service capacity.

A good theory, according to the system dynamics paradigm, links observable "macro", i.e., system-wide, patterns of behavior to "micro" decision making. Our 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 worked with us 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 (1990a, p 216-217):

"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." Because they are easier to measure, the team felt that there were natural pressures to manage by the production measures. As the vice president put it, "In this business there are lots of ways to look good without being good."

Initial simulation tests of the service quality-service capacity model showed how focusing on production standards could be problematic under times of stress. 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. 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 going on at a deeper level: desired levels may be

---

\(^1\) Complementary accounts of this study can be found in (Kim, 1991; Senge, 1990a; Senge, 1990b; Senge & Lannon, 1990; Senge & Sterman, 1992)

SYSTEM DYNAMICS '93 477
maintained only because of eroding quality. This model behavior reflected the vice president’s statement that it is “easy to look good without being good.” 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 behavior pattern which matches qualitatively the historical pattern of falling expense ratios and rising loss ratios (Senge, 1990a, pp. 220-227).

The basic formulation for the insurance industry is applicable to a set of service settings where providing the service involves highly intangible actions like building trust with the customer, where professional skills and considerable experience are required to successfully provide the service, and where complex personal interactions between server and customer are involved. In the four years since the original theory of capacity-quality-cost interactions was developed for insurance application, the basic model has been used in workshops for hundreds of managers from diverse service industries. This has led to a recasting of the original model as a generic theory. The full theory can be summarized by the following propositions:

1. In service businesses it is always difficult to measure quality because it is intangible and subjective.
2. Consequently, there is a tendency to manage service businesses by what is more measurable—notably, expenses and production figures.
3. This leads to a systematic bias toward underinvestment in “service capacity”—the ability to provide services at a given quality level, which is a function of number of people, experience levels, skills, and supporting infrastructure. Decisionmakers tend to assess whether or not capacity is adequate based on expenses and production figures, which may be unrelated to service quality.
4. The consequence of underinvestment is 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.
5. Underinvestment in service capacity is frequently masked by eroding operating standards, so that serving people and customers come to expect mediocre service and justify current performance based on past performance, rather than on absolute standards or goals.
6. As entire industries become locked in cycles of underinvestment and eroding standards, industry norms reinforcing expense control and “productivity” become more and more influential in shaping individual firm decisions.

The generic theory of service quality and service capacity has been elaborated in a system dynamics model. An overview of the model and the essential feedback relationships expressing the theory are presented below.

**SYSTEM DYNAMICS ARTICULATION OF THE GENERIC THEORY**

The Service Quality/Service Capacity model simulates a “Service Center” where customers enter the system and, after a waiting-time, are served by the Center’s employees. Service capacity — i.e., service personnel, years of experience, skill and motivation — is required to provide that service; the desired amount of capacity is determined by the desired level of quality, and the desired throughput of the Service Center. If a particular request is not satisfied to the customers’ standards, it comes back into the Service Backlog and has to be reprocessed as rework.

To understand the explicit formulations of the Service Quality/Service Capacity model, it is best to differentiate four subsystems in the model2: Service Capacity, Service

---

2 The following presentation is based on the “Service Quality Management. Flight Simulator Facilitators’ Training Guide”, prepared by Rogelio Oliva (1992), available from the MIT Organizational Learning Center.
Backlog, Quality/Time Pressure and Market Response. Figure No. 1 shows these subsystems and the main relationships among them. Each of the subsystems will be explained with more detail below.

**Service Capacity Subsystem**

The capacity sector assumes a chain of experience structure for the development of capacity. Personnel hired recently (Rookies) will not be as effective as Sr. Personnel until they go through a training and experience-building period (see Fig. No. 2). Furthermore, the training of Rookies will demand some time from the Sr. Personnel, reducing further the total capacity of the system.

Personnel turnover is determined by a normal turnover rate that is further modified by two factors: a) a Burnout Index that is the accumulation of stress due to the work intensity (Homer, 1985), and b) the Employees Perception of Service Quality. The Human Resources literature has extensive research that shows that employees will support more pressure and develop greater loyalty to the organization if they perceive a high service quality (Schneider, 1991; Schneider, Parkington, & Buxton, 1980; Tornow, 1991). The relationships between these indicators can be seen in Fig. No 3.

The current version of the simulator does not include the impact of macroeconomic indicators, i.e., unemployment, in the turnover rate. In the simulator, the Hiring rate is left as a decision for the user of the simulator although it can be set to keep a constant head-count. In the full simulation model, hiring rates are determined by customer flows and work backlog pressures (Senge 1990a).

**Service Backlog Subsystem**

The Service Backlog Subsystem keeps track of all the customer orders as they flow through the Service Center. Two stocks—Service Backlog and Undiscovered Rework—are used to model those dynamics. Figure No. 4 shows the main variables of this sector and their relationship.

One of the strongest assumptions in this particular model is that to deliver quality more time is required from the service provider. This implies that the Work Completed is reduced if the Actual Quality is increased. This was clearly the case for the claims adjusters because in order to do a better investigation, and keep more complete and accurate records, they had to spend more time with each claim. It is generally the case in service settings where customer interactions are important and perceived service quality suffers if customers feel rushed. In such settings, even though technology may improve efficiency in some aspects of the service provision, it cannot substitute for human contact. The implications of this assumption are going to be explored in the next section.

According to this diagram, the greater the Time Pressure, the lower the Actual Quality —employees will not have time to inspect their work or to respond to specific queries. This, assuming a constant work completed, will cause a higher rate of Mistakes that, after a period, are discovered and come back to the system as Rework, increasing the Service Backlog. A higher Service Backlog will be translated
into higher Production Goal as management seeks to reduce backlogs, and higher Time Pressure, thus reinforcing poorer quality. The way the system is formulated the balancing loop from Production Goal to Work Completed is always stronger than the reinforcing loop. However, 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).

**Quality/Time Pressure Subsystem**

This subsystem does not represent a physical department of the Service Center, nor a specific function of it. The Quality/Time Pressure Subsystem captures the tradeoffs and relationships between quality and time pressure. The speed at which a quality improvement process will achieve its objectives, and the impacts on quality of an increase of demand will be determined by this subsystem. To capture the dynamics of the service quality, it was necessary to define three different indicators for it.

**Quality Goal:** Reflects the management's desired level of quality. The Quality Goal is translated into pressure to modify the Quality Standard.

**Quality Standard:** The current Quality Standard is the employees' and management's perception of what is acceptable service quality, i.e. the quality level that the employees would perform under normal Time Pressure and Work Intensity.

**Actual Quality:** Actual Quality is the level of quality that the customers are receiving. The Actual Quality is based on the Quality Standard and Time Pressure.

The balancing loop in Figure No. 5 represents the relationships between quality and workload. The higher the Quality Standard is set, the more Effective Time Required to perform the job (more inspections, more quality improvement activities or more time spent in each transaction). This will, assuming a constant Productive Time Available, increase the Time Pressure. Because of hastedness induced by the Time Pressure, errors could be expected thus having a negative impact on the Actual Quality achieved. If the Actual Quality is lower than the Quality Standard for a long period, the standard then will begin to deteriorate as employees 'get used' to the lower quality level.

The reinforcing loop between Quality Standard and Actual Quality has different strengths depending on whether it is functioning as a vicious or a virtuous cycle. If the Actual Quality is lower than the Quality Standard and there is no pressure from the Quality Goal, the Quality Standard will begin to drift downward causing the Actual Quality to drop—even under constant Time Pressure. On the other hand,
if the *Quality Standard* is set at a value where the *Time Pressure* is less than normal, the *Actual Quality* achieved will be higher than the *Quality Standard*. This does not mean that the *Quality Standard* will increase with the *Actual Quality*, unless, of course, there is some pressure from the *Quality Goal*.

<table>
<thead>
<tr>
<th>Internal variable</th>
<th>Market Indicator</th>
<th>Perception Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Backlog</td>
<td>Average Waiting Time</td>
<td>3</td>
</tr>
<tr>
<td>Actual Quality</td>
<td>Customers' perception of Service Quality</td>
<td>6</td>
</tr>
<tr>
<td>Time Pressure</td>
<td>Sales Effort from the Service Personnel</td>
<td>1</td>
</tr>
</tbody>
</table>

**Market Response Subsystem**

Table No. 1

The *Relative Attractiveness* of the service provided by the Service Center is determined by three market indicators that are linked, through perception delays and biases, to internal variables of the Service Center (see Table No. 1). All these indicators form negative loops regulating the *Customer Orders* through the *Relative Attractiveness* of the service.

**COMPARISON TO OTHER MODELS AND THEORIES OF SERVICE QUALITY**

As the Service sector is becoming increasingly important in the US economy (Cohen & Zysman, 1987; Quinn & Gagnon, 1986), more time has been spend in trying to develop guidelines to manage services. In this section we will identify how our dynamic theory relates to other emerging service quality models and theories.

Much of the research effort on the service industry, especially in the marketing and operations management literature, has focused in determining the nature of service and identifying the different kinds of settings under which services are provided. There are many classifications of services available (Haywood-Farmer, 1988; Lovelock, 1983; Schmenner, 1986), each focusing in additional distinctions in order to define strategic or operational guidelines for the service managers. It is difficult, however, to build a consistent view of service quality from all those models and classifications. Perhaps, the important lesson from this proliferation of service models is that the determinants of service quality will vary with the type of business studied. The different classifications of services, are however, useful to explore the limitations and shortcomings of the model presented here.

a) As discussed above, the current version of the model assumes that an increase in *Actual Quality* will necessarily imply a lower rate of *Work Completed*. It is, however, difficult to make the generalization of this relationship since one of the strongest argument in the Total Quality Management literature is that total cost is reduced through better quality (Crosby, 1979; Deming, 1982; Juran, Gyna, & Bingham, 1974). The model, as it stands, does reflect some reduction of total cost if *Actual Quality* is improved—through reduced *Rework* and lower *Personnel Turnover*. The model, however, is not capable of capturing the increase on productivity through the standardization process that might emerge after a technological or process improvement of quality.

b) The current formulation of the theory takes the only indicator of capacity to be personnel and their effectiveness. It is not possible to capture other 'capacity increasing' investments. This particular conceptualization of capacity limits the learning curve to enhanced productivity through work experience.

Although formulations to deal with those limitations can easily be incorporated into the model, the fact that we have not done it limits the settings in which the model can be used. The model is currently best suited to represent service centers where the service interaction is complex, the average waiting time is in the order of days or weeks, not hours or minutes, and the service provided depends on highly trained personnel.

Regardless of the diversity of service settings, there are a core of concepts widely accepted as intrinsic to services (Chase, 1981). Among those concepts is the recognition that services are intangible activities, thus the difficulty of measuring their quality. This recognition of intangibility has brought up the recognition that service quality perceptions result from a comparison of customer expectations with
service performance (Maister, 1984). The most articulated model of this perspective (Parasuraman, Zeithaml, & Berry, 1985) argues that the difference between the customer expectations and the actual service provided cannot be managed directly but through other “gaps,” or discrepancies, between expectations and performance that occur in organizations. Figure No. 6 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 (*Quality Goal* in our model) and actual service quality specifications (*Quality Standard*);

**Gap 3** the difference between service quality specifications and the service actually delivered (*Actual Quality*);

**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.

Our model, as it stands, assumes perfect knowledge of the customer and only explores the dynamics of the internal organization—Gaps 2 and 3 (shaded area in Figure No. 6). We believe that these dynamics are taken-for-granted in most of the Total Quality Management literature and it is worth exploring them because they have first order effects in practical managerial settings.

We believe that our theory, although limited in the scope of total quality, addresses significant interdependencies that constitute a meaningful "internal perspective" on service quality.

In relation to other studies of service quality, this theory has certain distinct advantages, namely

1. Endogenous theory of dynamics: the theory, when simulated, recreates a pattern of underinvestment, eroding quality, increasing costs of poor quality and deteriorating overall financial performance that occurs in many real service industries.

2. All assumptions are explicit: true for other mathematical models of services, but most of these are not fully dynamic.

**STRATEGY FOR BUILDING CONFIDENCE IN THE THEORY**

The generic theory of service quality-service capacity interactions is in a formative stage. The insurance case provided one context for development and testing. The results were sufficiently encouraging to pursue other contexts. In those settings we will endeavor to build on and go further in testing the theory than was feasible in the first setting.

Understanding a system dynamics theory requires an intuitive appreciation of "its dynamics." The model is an artifact that expresses a theory in an interactive form. The theory can therefore only be fully understood through interacting with it. This is why actual simulation experience is so important to
the validation process. Building confidence in the theory is a continual process of deepening understanding of its dynamics and checking against experience in the system being studied, often against the experiences of people who live in the system.

This is why the "validation process" cannot be reduced to a few simple tests or summary presentations of how the model behaves. Rather it is a never-ending process of performing increasingly diverse tests (Forrester & Senge, 1980; Sterman, 1988). For example, for the claims management simulator described above, there are basic "model behavior tests" that seek to determine that the model can reproduce salient historical behavior patterns like rising loss ratios, without the aid of external inputs that might predetermine such results. As illustrated, the original model passed such tests. But, validation also requires "policy testing" to ascertain how the model responds to different changes in management policies. For example, total costs in the model do not improve simply by increasing quality goals. In fact, elevating quality goals alone can be counterproductive, leading to overwork, burnout, increased turnover and increasing delays in customer service.

The basic strategy for continuing to test the generic service capacity-service quality theory is to use it as a basis for theory development and intervention in a variety of service companies participating in the Center for Organizational Learning. In each case, the basic precepts and interdependencies incorporated into the theory will be held up against the specific world of a particular group of managers in a particular industry. Developing learning laboratories like that developed in the original insurance study will guarantee that the theory will come in contact with a large number of managers. Hopefully, it will also stimulate a variety of organizational experiments based upon its implications, which will provide further data regarding the model's usefulness and reveal limitations and flaws that can lead to further improvement.

The general approach that we have decided to take is to explore the 'range of usefulness of the model.' This might be done by testing the overall behavior and implications with the managers, and verify the transferability of the model (assumptions) through several different service settings in an empirical way. The overall objectives of the empirical testing are to

1. gain more experience in the practical utilization of the theory
2. gain better understanding of the theory's strengths and limitations
3. provide stimulus to develop new and better theories and better practical tools.

The range of activities now planned are summarized below. As with all validation strategies, each particular type of test has its own limitations, which are also briefly summarized.

1. Organize conceptualizing sessions where managers can articulate their perceptions of service-quality-service capacity interactions in feedback terms that can be related to the theory.
   *Limitation:* can get a wide variety of inputs based on very different mental models; commensurability with core theory can be challenging'

2. Collect data in companies that might assist in further testing of overall patterns of model behavior.
   *Limitation:* available numerical data are often limited

3. Develop Learning Laboratory in each company, observe decision making behavior within these labs, and relate to decision rules incorporated into full simulation model.
   *Limitation:* do people behave in experimental setting in a way that is representative of their behavior in real decision making settings?

4. Test policy implications of theory through working with managers to translate implications, implement changes, and study the consequences of those changes.
   *Limitation:* messy "data" that comes from interventions and assessing the consequences of interventions in the presence of confounding factors and long delays

Ultimately, this leads to an approach to validation which puts theory testing within the context of organizational learning. The roots of the term "valid" are in law and have to do with the logical consistency and usefulness of an argument, and not contradicting known facts. Thus, it makes very good sense to pursue a validation strategy of having managers interact with a model, deepen their intuition about its dynamics, draw conclusions about possible implications for their management.
practice, undertake experiments to test out those conclusions in the real system, and feed back new insights that enrich the basic theory. Kofman (1992) has articulated the overall organizational learning process in terms of observation, assessment, design and implementation or intervention. We believe that this process can also serve to test theories guiding that learning.

In this context, the system dynamics theory is a tool to foster new, more insightful "observations" based on a more systemic interpretation of the "data of experience." It is also to be used in assessing likely effects of alternative actions that might be taken to improve system performance and in designing specific interventions. These actions usually take the form of new organizational structures, rewards, performance measures or other changes in operating policies. The merits of the theory are then ascertained through studying the effects of an actual implementation of new policies.

To illustrate, the original claims management learning laboratory was not intended to dictate particular changes that managers should make but rather to raise important questions about interdependencies among costs, quality, and production. The intention was to send managers back into their field operations strategizing about how they might improve total costs. For example, one participant reported:

"When I came back from the learning laboratory, I had a much better understanding of what the important issues were. Before the lab, I would have said that lack of quality was the only important factor. After the lab, it was obvious to me that productivity was also a key issue. So I restructured some units to enhance their ability to settle claims. After I saw dramatic increases in productivity [in the real organization], I applied pressure to improve quality—and I have seen a difference." (Bergin & Prusco, 1990).

REFERENCES


