A Psychological Perspective on Service Segmentation Models: The Significance of Accounting for Consumers' Perceptions of Waiting and Service

Ziv Carmon • J. George Shanthikumar • Tali F. Carmon
Marketing Group, Fuqua School of Business, Duke University, Durham, North Carolina 27706
Management Science Group, Walter A. Haas School of Business, University of California, Berkeley, California 94720
Department of Industrial Engineering and Operations Research, University of California, Berkeley, California 94720

We examine how service should be divided and scheduled when it can be provided in multiple separate segments. We analyze variants of this problem using a model with a conventional function describing the waiting cost, that is modified to account for some aspects of the psychological cost of waiting in line. We show that consideration of the psychological cost can result in prescriptions that are inconsistent with the common wisdom of queuing theorists derived according to the conventional approach (e.g., equal load assignments). More generally, our intention in this paper is to illustrate that aspects of the psychological cost of waiting can be accounted for in the analysis of queuing systems, and that this may have significant implications for the service schemes that are derived.

(Queuing; Psychology of Waiting; Dissatisfaction; Workload Allocation; Service Ordering)

1. Introduction

People typically do not like to wait, as it causes them to experience a broad range of unpleasant responses such as boredom, irritation, anxiety, tension, helplessness, and sometimes even humiliation. Yet waiting for service is one of our most common daily experiences. Banks, restaurants, public transportation, medical treatment, and operator assistance are just a few examples of service establishments in which consumers are required to wait. Accordingly, due to the phenomenal growth of the service sector in the American economy and the increasing competition in many of the service markets, consumers' satisfaction with service has become a major concern for academics and practitioners alike (Lindley 1988). Management of consumer queuing systems should therefore presumably be an extensively researched area. Surprisingly, this is true in only a limited sense.

To date, much of the research on queuing has been carried out by operations researchers concerned with mathematical modeling of waiting lines. Numerous published studies have successfully explored characteristics of queuing systems in a wide variety of contexts to provide information for effective decisions about service policies. However, for consumer service environments in which the cost of waiting is endured by people who wish to be served, conventional queuing research has not been as effective. Although Queuing Theory considers the obvious costs of waiting in terms of time and money, it does not account for the psychological cost. Marketing practitioners have long known that the psychological cost, which is commonly referred to as consumer dissatisfaction, has a significant effect on subsequent behavior, such as the likelihood of future visits to the service establishment. This link between satisfaction and succeeding actions has also been doc-
umented by marketing researchers (Boulding et al. 1993, Anderson and Sullivan 1993). Note that throughout this paper, we loosely refer to the consequences of waiting mentioned in the first paragraph as psychological cost, dissatisfaction, or disutility.

In this paper we study a classic problem in queuing research, and demonstrate that an analysis which considers consumer satisfaction can result in service schemes that are significantly different from ones based on the traditional analysis. Specifically, we examine how service should be divided when it can be provided in multiple separate segments, and how such service (that is already allocated into given segments) should be scheduled. For example, in commercial flights, the primary service is flying customers to their destination, but "filler" services include serving snacks, drinks, and meals, and screening movies in the course of the flight. These fillers do not affect the time customers spend in the system, but are instrumental in shifting attention away from the passage of time so that the customers do not become preoccupied with their discomfort. The fillers can be sequenced and scheduled in a variety of ways, each of which may affect consumers' dissatisfaction differently. A similar problem can arise in fast-food restaurants, where order-taking, payment, serving items that are immediately available (such as soft drinks), and serving the items that must be prepared, can be segmented and scheduled in several different ways. Note that in the latter example, the service scheme may affect both the total duration and customers' dissatisfaction. The choice of the specific problem we study in this paper was largely motivated by its prominence in the classical queuing literature (see e.g., Buzacott and Shanthikumar 1992).

We construct a simplified model of the cost of waiting that focuses on the effects of the passage of time, of providing service and of its timing. Beyond mathematical tractability considerations, this decision was guided by our desire to keep the model simple and as similar as possible to conventional queuing frameworks. The choice of this focus was also motivated by recent research which revealed that in people's evaluations of extended experiences, duration has a small, and at times, an insignificant effect (Fredrickson and Kahneman 1993). Instead, structural aspects of a situation, such as the timing of events within a given sequence, are consequential (see, e.g., Loewenstein and Prelec 1993, and Carmon and Kahneman 1993). These findings imply that the timing of different phases of a service may affect consumers' overall (dis)satisfaction. This aspect of waiting experiences is not reflected in traditional queuing modeling due to the focus on physical performance measures, particularly the total time spent in the queue.

We investigate two types of conditions. We first study situations in which it can be reasonably assumed that the sojourn time (total time spent in the system) is not affected by the service scheme (i.e., the sojourn time can be assumed to be constant). This is true, for example, in the commercial flights example discussed above. We then explore cases in which the service scheme may affect the sojourn time (i.e., the sojourn time cannot be assumed to be constant), as is true of the fast-food restaurant example we described. However, since this latter analysis is considerably more complex, we focus on a specific example that allows us to illustrate inconsistencies both with traditional solutions and with those we derive for comparable cases in which the sojourn time is fixed.

The remainder of this paper is organized as follows. Since our analysis requires assumptions about the cost of waiting and the benefits of service, in §2 we discuss the characteristics of the cost function and the effect of service on this cost (the service function). We also consider the choice of measures of the disutility associated with waiting, and describe our model and some of its basic characteristics. In §3 we present the analysis of the problem, and contrast prescriptions that are based on our model with those that would be derived according to conventional queuing research. §4 consists of concluding remarks regarding our findings, their implications, and suggestions for future research.

2. An Analytical Representation of Dissatisfaction

2.1. Assumptions

Studies of people's responses to a variety of aversive sensations, such as pain (Wolff and Wolf 1958), living near a noisy highway (Weinstein 1982) and waiting, (Palm 1953, Osuna 1985) have revealed that these
responses escalate over time. In particular, when the expected duration cannot be estimated (e.g., when the cause of the delay is unknown), it has been suggested that the disutility associated with waiting increases with time (Osuna 1985). Service, on the other hand, generally reduces consumers' dissatisfaction. This is because customers are provided with benefits that compensate for the perceived cost of waiting, and since attention is attracted away from the passage of "empty" time (Fraisse 1963).

We also assume that the rate at which dissatisfaction changes over time depends only on the level of dissatisfaction at that time. Hence, we are assuming that the effect of one's past in the queue can be captured by a single parameter, namely, the current level of dissatisfaction. This assumption, which is required for mathematical simplicity, cannot be directly supported by empirical evidence since this specific issue has not been investigated. However, as a first approximation, it seems reasonable and is consistent with the work of several researchers in the field of psychology (e.g., Helson 1964, Linville 1987, Linville and Fischer 1991). Note that some idiosyncrasies of the system can be captured by the functional form and the parameters of the waiting and service functions.

2.2. Measures of the Psychological Cost

Two candidates for a measure of the psychological cost reflect two potential consequences of waiting for service. If consumers' dissatisfaction in the course of waiting rises beyond a threshold level, they are likely to balk (leave without receiving service). Alternatively, if they feel dissatisfied after service is completed, they will be less likely to return to the establishment, and be more likely to spread negative word-of-mouth that could affect the likelihood of other consumers' visits. Hence, the respective dependent variables are the maximal in-process dissatisfaction and the dissatisfaction once the service is completed. In this paper we choose the latter, emphasizing the long-term effects of excessive waits, since we believe that they can be easily overlooked. Moreover, consumers' final dissatisfaction is critical in what we name necessary services in which customers cannot balk (e.g., monopolistic or urgently needed services).

Specifically, the paper focuses on customers' dissatisfaction at the time of departure from the service. Beyond considerations of mathematical tractability, this decision is motivated by recent findings of Carmon and Kahneman (1993), who have shown that consumers' final dissatisfaction with waiting for service is very highly correlated with their global retrospective (dis)satisfaction judgments. These judgments then have an effect on subsequent actions (Boulding et al. 1993, Anderson and Sullivan 1993).

2.3. Basic Characteristics of the Model

Let \( r_w(x) \) be the rate at which dissatisfaction changes during waiting, and \( r_s(x) \) be the rate at which it changes during service when the level of dissatisfaction is \( x \). We will assume that \( r_w(x) \) is positive (i.e., the level of dissatisfaction increases during waiting), finite for all \( x < \infty \), and increasing in \( x \). Throughout this paper, 'increasing' and 'decreasing' are not used in a strict sense. We also assume that \( r_s(x) \) is finite for all \( x < \infty \). Note that \( r_s(x) \) could be positive for some values of \( x \) and negative for other values, indicating that the level of dissatisfaction could decrease during service.

Let \( g(t) \) be the dissatisfaction level at time \( t \). We will assume that \( g(t) \) is continuous in \( t \). Suppose the consumer is waiting at that time (i.e., not receiving service at time \( t \)). The function \( r_w \) is then formally defined by

\[
r_w(g(t)) = \lim_{\tau \to t} \frac{g(\tau) - g(t)}{\tau - t}.
\]

Let \( h(t) \) be the level of dissatisfaction at time \( t \) if service is not provided during \([0, t]\). Therefore, if the consumer's initial dissatisfaction level is \( h(0) \), the dissatisfaction at time \( t \) is described by the following integral equation:

\[
h(t) = h(0) + \int_0^t r_w(h(\tau))d\tau, \quad t \geq 0.
\]

Remark 1. \( h(t) \) is found through solving the above integral equation. Since \( r_w(x) \) is positive, \( h(t) \) is an increasing function of \( t \). For example, if \( r_w(x) = c \) is a constant independent of the level of dissatisfaction, we get \( h(t) = h(0) + ct, \ t \geq 0 \), while if \( r_w(x) = cx \), that is, linearly increasing with the level of dissatisfaction, we get \( h(t) = h(0)e^{ct}, \ t \geq 0 \).

We will now formally define \( r_s(x) \). Let \( g(t) \) be the level of dissatisfaction at time \( t \). If the consumer is receiving service at time \( t \), then
Let \( f(t) \) be the level of dissatisfaction at time \( t \) if the customer receives service during \([0, t]\). Therefore, if \( f(0) \) represents the consumer’s initial dissatisfaction level, the dissatisfaction at time \( t \) is described by the following integral equation:

\[
f(t) = f(0) + \int_0^t r_s(f(\tau))d\tau, \quad t \geq 0.
\]

(4)

For example, if \( r_s(x) = -x \), that is, linearly decreasing with the level of dissatisfaction, we get \( f(t) = f(0)e^{-t}, t \geq 0 \).

Let \( G(w, s) \) represent the dissatisfaction level of a consumer who waits for \( w \) units of time before receiving a service of length \( s \). Then from (2) and (4) one sees that the level of dissatisfaction \( G(w, s) \) at time \( w + s \) is given by \( g(w + s) \), where:

\[
g(t) = g(0) + \int_0^t r_s(g(\tau))d\tau, \quad 0 \leq t \leq w,
\]

(5)

\[
g(t) = g(w) + \int_w^t r_s(g(\tau))d\tau, \quad w \leq t \leq w + s.
\]

(6)

Recall that \( h(t) \) is the level of dissatisfaction when a consumer spends \( t \) units of time in the system without receiving service (see Equation (2)). Let \( \delta \) be selected such that:

\[
g(w + s) = h(w + \delta).
\]

(7)

From (7) we see that the consumer’s level of dissatisfaction at time \( w + s \), after receiving service for \( s \) units of time, is the same as that of a consumer who waits for \( w + \delta \) units of time without receiving service. According to this perspective, we may view the effect of service as an adjustment of the effective waiting time. Thus, the effect \( \delta \) of service may be conceptualized as reducing the effective waiting duration. Note that by the effective waiting time, we refer to the duration \( (w + \delta) \) of waiting (without service) that has an effect on dissatisfaction that is equivalent to that of the combination of the waiting and the service times experienced. This reduction in effective waiting due to service will happen when \( r_s \) is a negative function. Note, in contrast, that the objective function in the traditional queuing analysis simply treats the service as additional waiting time from the consumer’s view (i.e., the sojourn time \( w + s \) is the measure of waiting in the system).

Let \( f(s, t) \) be the portion of the effective waiting time accumulated during the service period if the service is initiated when the level of dissatisfaction is \( h(t) \) and is continued for \( s \) units of time. Hence, the level of dissatisfaction at the termination of the \( s \) units of service is \( h(t + f(s, t)) \). From (3) observe that when \( h(t) = x \), then \( dt/dx = 1/r_s(x) \). Thus, the rate of passage of time with respect to the level of dissatisfaction during service is \( 1/r_s(x) \) when the level of dissatisfaction is \( x \). Since \( s \) is the duration of change in the level of dissatisfaction from \( h(t) \) to \( h(t + f(s, t)) \), we have:

\[
s = \int_{h(t)}^{h(t + f(s, t))} \frac{1}{r_s(x)} dx.
\]

(8)

Taking the derivative of both sides of Equation (8) with respect to \( s \), one gets:

\[
\frac{d}{ds} f(s, t) = \frac{r_s(h(t + f(s, t)))}{r_w(h(t + f(s, t)))}.
\]

(9)

From Remark 1 and Equation (9), the following is then immediate:

**LEMMA 1:** If \( r_s(x)/r_w(x) \) is decreasing (increasing) in \( x \), then \( f(s, t) \) is decreasing (increasing) in \( t \).

**PROOF:** Define \( \gamma(x) = [r_s(h(t + x))]/[r_w(h(t + x))] \). From Remark 1 one sees that if \( r_s(x)/r_w(x) \) is decreasing (increasing) in \( x \), then \( \gamma(x) \) is decreasing (increasing) in \( x \) and \( t \). Rewriting Equation (9) one has:

\[
\frac{d}{ds} f(s, t) = \gamma_t(f(s, t)).
\]

(10)

From (10) and Lemma A1 (given in the Appendix) it can be seen that for \( t_1 < t_2 \), we have:

\[
f(s, t_1) \geq (\leq) f(s, t_2).
\]

(11)

This completes the proof.

**REMARK 2.** The above lemma states that if the ratio of the service function’s rate to the waiting function’s rate decreases with the level of dissatisfaction, the effectiveness of service increases with increasing dissatisfaction. On the other hand, if this ratio increases, then the effectiveness of service decreases with increasing dissatisfaction. Thus, based on this lemma, we can determine the timing for providing service as a function...
of this ratio, so as to minimize consumers' dissatisfaction at the end of the service encounter. These conclusions will be formally introduced in Theorems 1 and 2.

**Remark 3.** Note that if \( r_t(x)/r_w(x) \) is independent of \( x \), then \( f(s, t) \) is independent of \( t \). In other words, the effectiveness of service is independent of the level of dissatisfaction.

We now explore the effect of service that is provided in the course of a waiting period. Let \( G(w_1, s, w_2) \) be the level of dissatisfaction at time \( w_1 + s + w_2 \) if a consumer waits for \( w_1 \) units of time, receives service during \([w_1, w_1 + s]\) and waits for another \( w_2 \) units of time before leaving the system. Therefore,

\[
G(w_1, s, w_2) = h(w_1 + f(s, w_1) + w_2). \tag{12}
\]

The following lemma will be used to decide whether the service should be offered at an early stage of the waiting period or at a late one.

**Lemma 2.** For a fixed value of \( s \), if \( r_t(x)/r_w(x) \) is decreasing (increasing) in \( x \), then \( G(w_1, s, w_2) \geq (\leq) G(w'_1, s, w'_2) \) whenever \( w_1 \leq w'_1 \) and \( w_1 + w_2 = w'_1 + w'_2 \).

**Proof.** From Lemma 1 we know that if \( r_t(x)/r_w(x) \) is decreasing (increasing) in \( x \), then \( f(s, w_1) \) is decreasing (increasing) in \( w_1 \), and hence for \( w'_1 \geq w_1 \) one obtains \( f(s, w'_1) \leq (\geq) f(s, w_1) \). Combined with \( w'_1 + w'_2 = w_1 + w_2 \), derived from the increasing property of \( h \) (see Remark 1) and Equation (12), we get

\[
G(w'_1, s, w'_2) = h(w'_1 + f(s, w'_1) + w'_2) \\
\leq (\geq) h(w_1 + f(s, w_1) + w_2) \\
= G(w_1, s, w_2). \tag{13}
\]

This completes the proof.

### 3. Traditional vs. Psychologically Based Queuing Analysis

In §2 we illustrated the feasibility of constructing an analytical model that accounts for some aspects of the psychological cost of waiting. In this section, we demonstrate that prescriptions that are based on this model can be inconsistent with those of traditional queuing analysis. For that purpose, we derive service policies (optimal segmentation of a service, and scheduling service segments) for two types of conditions, ones in which the sojourn time can be assumed to be unaffected by the service policy and ones in which this cannot be assumed. Recall that our objective is to minimize the dissatisfaction level at the end of the service encounter since Carmon and Kahneman (1993) have shown it to be highly correlated with the global retrospective dissatisfaction, which affects consumers' future actions.

#### 3.1. Fixed Sojourn Time

Traditional analysis typically seeks to minimize the sojourn time. From this perspective, in the first type of problems (constant sojourn time) the prescription is that the service scheme does not matter, since the service policy does not affect duration. In some situations our analysis will also result in this conclusion. For example, consider a case in which the level of dissatisfaction remains constant while service is being provided, and, once it is terminated, dissatisfaction continues to increase at the same rate as before service was initiated. In this situation, it is clear that the service scheme does not matter. This is because providing service, effectively, merely shortens the overall waiting duration by a fixed magnitude and has no effect on the rate at which dissatisfaction increases over time.

As another example, consider the case \( r_w(x) = -cr_t(x) \). This is a context in which the rate of increase in dissatisfaction while waiting is linearly related to the rate at which dissatisfaction drops when service is provided. If, for example, the two rates are equal, providing service of a given duration can be thought of as an elimination of a waiting duration of equal length (recall that, according to our assumptions, both rates depend on the momentary level of dissatisfaction). If that is the case, then it obviously does not matter when this elimination occurs. Similarly, if the rates are proportional, providing service is equivalent to the elimination of a waiting period for which the duration is equal to the service duration times the absolute value of the ratio of the service rate to the waiting rate. Thus, in such cases it does not matter when service is given, as is suggested by conventional analysis. However, there are many contexts in which our prescriptions will deviate from those derived according to the traditional approach, as we illustrate in the analysis and the examples to follow.
One of the main results for the service scheme in the fixed sojourn time case follows immediately from Lemma 2. When the ratio of the service function’s rate of change to the waiting function’s rate of change is decreasing with the level of dissatisfaction, the relative effectiveness of service increases with increasing dissatisfaction, and service should therefore be provided as late as possible. However, if this ratio is increasing, then the same reasoning leads to the opposite conclusion. This prescription can be applied in situations such as the commercial flight example described in the introduction. For instance, if there is a movie to be screened, according to our analysis it should either happen at the very beginning of the flight or toward its end. This logic is formally presented in Theorems 1 and 2.

**THEOREM 1.** Consider providing a single service (say of a fixed duration $s$) in a situation in which the sojourn time is fixed. If $r_1(x)/r_w(x)$ is decreasing (increasing) in $x$, it is better to provide service later (earlier) than earlier.

A similar analysis leads to the following:

**THEOREM 2.** When the sojourn time is fixed, consider providing two service portions (of fixed durations $s_1$ and $s_2$, $s_1 < s_2$) after waiting periods for which the durations are fixed. If $r_1(x)/r_w(x)$ is decreasing (increasing) in $x$, it is better to provide the shorter (longer) service first.

**PROOF.** Let $w_1$ and $w_2$ be the waiting times before the first and the second service segments. The level of dissatisfaction at time $w_1 + s_1 + w_2 + s_2$ (see Equations (5), (6) and (12)) is given by:

$$G(w_1, s_1, w_2, s_2) = h(w_1 + f(s_1, w_1) + w_2 + f(s_2, w_1 + f(s_1, w_1) + w_2)) \tag{14}$$

From Equation (9) it can be observed that:

$$f(a + b, t) = f(a, t) + f(b, t + f(a, t)). \tag{15}$$

Therefore,

$$f(s_2, w_1 + f(s_1, w_1) + w_2) = f(s_2 - s_1, w_1 + f(s_1, w_1) + w_2) + f(s_1, w_1 + f(s_1, w_1) + w_2) + f(s_2 - s_1, w_1 + f(s_1, w_1) + w_2). \tag{16}$$

Since $f(s, t)$ is decreasing in $t$ (see Lemma 1), from (16) one sees that:

$$f(s_1, w_1) + f(s_2, w_1 + f(s_1, w_1) + w_2)$$

$$= f(s_1, w_1) + f(s_2 - s_1, w_1 + f(s_1, w_1) + w_2) + f(s_1, w_1 + f(s_1, w_1) + w_2)$$

$$+ f(s_2 - s_1, w_1 + f(s_1, w_1) + w_2) \geq f(s_1, w_1) + f(s_2 - s_1, w_1 + f(s_1, w_1))$$

$$+ f(s_1, w_1 + f(s_1, w_1) + w_2) + f(s_2 - s_1, w_1 + f(s_1, w_1)))$$

$$= f(s_2, w_1) + f(s_1, w_1 + f(s_2, w_1) + w_2). \tag{17}$$

Combining Equations (14) and (17) one sees, according to Remark 1, that:

$$G(w_1, s_1, w_2, s_2) = h(w_1 + f(s_1, w_1) + w_2$$

$$+ f(s_2, w_1 + f(s_1, w_1) + w_2))$$

$$< h(w_1 + f(s_2, w_1) + w_2$$

$$+ f(s_1, w_1 + f(s_2, w_1) + w_2))$$

$$= G(w_1, s_2, w_2, s_1). \tag{18}$$

This completes the proof.

Thus, the general condition we propose for minimizing dissatisfaction is that if the ratio of the service function’s rate to the waiting function’s rate is decreasing with the level of dissatisfaction, as much of the service as possible should be provided as late as possible. On the other hand, if it is increasing with dissatisfaction then as much of the service as possible should be provided as soon as possible. Recall that the intuition for this condition is that when this ratio decreases with the level of dissatisfaction, for example, the relative effectiveness of service increases with increasing dissatisfaction, and consequently service should be provided as late as possible.

As a simple example, consider a case in which providing service leads to a drop in dissatisfaction at a constant rate, i.e., $r_1(x) = -c$. Then $r_1(x)/r_w(x)$ is increasing in $x$. In this setting, it is clear that as much of the service
as possible should be provided as soon as possible. Due to the convexity of the dissatisfaction function, service that is provided at an early stage will prevent a steep increase later on, as well as reduce the level of dissatisfaction by a fixed magnitude. On the other hand, service that is provided toward the end of the waiting period will reduce the level of dissatisfaction by the same magnitude, but have little effect on subsequent buildup of dissatisfaction.

Note that our results are generally consistent with psychological research on impression formation, according to which overall impressions are affected primarily by either the initial impressions or the final ones (see e.g., Fiske and Taylor 1984). Furthermore, the order in which information items are presented has also been shown to affect the global impressions. While the reasons for these conclusions may be different than the ones we rely on, we find the convergence encouraging.

Remark 4. Note that the idea underlying both Theorem 1 and Theorem 2 can be generalized. If service is to be provided in \( n \) portions, then, if \( r_s(x)/r_w(x) \) is decreasing (increasing) in \( x \), it is better to provide as many of the portions as possible later (earlier) than earlier (later). Furthermore, it is best to provide the portions in ascending (descending) order of their durations. Both generalizations can be proven by induction.

### 3.2. Variable Sojourn Duration

In the second type of situation the sojourn time is not assumed to be constant. We consider a two-stage tandem queuing system, where service is provided in two segments. The optimal policy must consider two conflicting dimensions. One is the waiting duration, which, all other things being equal, should be minimized. Indeed, traditional analysis prescribes assigning an equal portion of the service (on average) to each server. Otherwise, the server assigned the largest portion of the service becomes a bottleneck and increases the average waiting time. This intuition can easily be rigorously substantiated (e.g., see Buzacott and Shanthikumar 1992). In manufacturing settings, for example, equal load assignments are the typical scheme, since the cost associated with the service policy is primarily affected by the sojourn time. The other dimension that the optimal policy must consider is (dis)satisfaction, which should also be minimized in consumer service environments. This will favor providing more service at an early stage of the waiting period in some cases, and in others providing more service toward the end. Both cases involve an unequal work load. Loosely speaking, it is unlikely that combining these two motives will lead to an equal load scheme, which would be the conventional prescription. Our analysis will indeed show that in consumer service systems, the equal load assignment need not be the optimal scheme. Moreover, we will demonstrate that the optimal service schemes in these contexts may also be inconsistent with ones derived for comparable situations in which the sojourn time is assumed to be constant.

We will now consider a stochastic example. Suppose that customers arrive at a two-stage tandem queuing system according to a Poisson process with a rate of 1. The service consists of two segments: one of length \( s_1 \) and the other of \( s_2 \). The actual service times will be assumed to be i.i.d. random variables, exponentially distributed with means \( s_1 \) and \( s_2 \), respectively. Furthermore, we will assume that the arrival process and the service times are all mutually independent. Suppose \( s_1 < s_2 \). The question is whether segment 1 of the service should be assigned to stage 1 or to stage 2 of the tandem queuing system. We will only consider an example of the case where \( r_s(x)/r_w(x) \) is decreasing in \( x \). Our objective here is to minimize the mean dissatisfaction of an arbitrary customer at the departure from the system.

**Example.** Suppose \( r_w(x) = 1; r_s(x) = -x \). Then \( h(t) = h(0) + t + f(s, t) = h(0) + t(1 - e^{-x}) \). From (14) it can be easily computed that when \( h(0) = 0 \),

\[
G(w_1, s_1, w_2, s_2) = w_1 e^{-w_1} e^{-s_2} + w_2 e^{-w_2}. \tag{19}
\]

Hence, if \( s_1 < s_2 \), then \( G(w_1, s_1, w_2, s_2) < G(w_1, s_2, w_2, s_1) \). That is, it is preferable to provide the shorter service first. This is not surprising, because in this case \( r_s(x)/r_w(x) \) is decreasing in \( x \) (see Theorem 2). In this analysis, however, we have \( w_1 \) and \( w_2 \) fixed, but in the queuing environment, the waiting times \( w_1 \) and \( w_2 \) are commonly random, and depend upon the allocation of the service times to stages one and two. Let \( W_i(a, b) \) be the stationary waiting time of an arbitrary customer in the \( k \)th stage of the tandem queue with a service time allocation of \( a \) units to stage 1 and \( b \) units to stage 2. Then \( E[W_i(a, b)] = a^2 / 1 - a \) and \( E[W_2(a, b)] = b^2 / 1 - b \) (see, for example, Kleinrock 1975). The average dissatisfaction
of an arbitrary customer for this allocation is then (see Equation (19))

\[
E[G(W_1(a, b), A, W_2(a, b), B)]
\]

\[
= E[W_1(a, b)]E[e^{-A}]E[e^{-B}]
+ E[W_2(a, b)]E[e^{-B}]
\]

\[
= \frac{a^2}{(1 - a^2)(1 + b)} + \frac{b^2}{1 - b^2}
\]

\[
= \frac{a^2}{1 - a^2} + \frac{b^2}{1 - b^2} - \frac{ab}{(1 + a)(1 + b)} \left( \frac{a}{1 - a} \right),
\]

(20)

where A and B are two independent exponential random variables with mean \( a \) and \( b \) respectively. It is then easily seen from (20) that for \( s_1 < s_2 \),

\[
E[G(W_1(s_1, s_2), S_1, W_2(s_1, s_2), S_2)] > E[G(W_1(s_2, s_1), S_2, W_2(s_2, s_1), S_1)].
\]

Therefore it is preferable to provide the longer service first. Observe that this contradicts the fixed sojourn time case, and both cases are inconsistent with the prescription of the traditional approach. The implication is that in a queuing environment one cannot rely on observations that were made for non-queuing situations in which the sojourn time is not affected by the service policy.

Now consider the same example, but suppose that we are interested in allocating a total service time of, say, \( s \) units among the two stages. We wish to obtain the optimal allocation that minimizes the mean level of dissatisfaction. Thus we have the problem:

\[
\min \{ E[G(W_1(a, b), A, W_2(a, b), B)] : a + b = s; a, b \geq 0 \}. \quad (P)
\]

From Equation (20) one sees that:

\[
\frac{d}{da} E[G(W_1(a, b), A, W_2(a, b), B)] |_{a=b=s/2}
\]

\[
= \frac{4s^2}{(4 - s^2)^2} < 0.
\]

That is:

\[
\frac{d}{da} E[G(W_1(a, b), A, W_2(a, b), B)] |_{a=b=s/2}
\]

\[
< \frac{d}{db} E[G(W_1(a, b), A, W_2(a, b), B)] |_{a=b=s/2}.
\]

Therefore, it is clear that the optimal solution to problem (P) is not \( a = b = s/2 \). Indeed, it can be verified that the optimal solution satisfies \( 1 > a > s/2 > b \) for all \( 0 < s < 2 \). This too is in contradiction to the traditional wisdom of balancing the workloads so as to minimize the mean sojourn time.

4. Concluding Remarks

There are two types of arguments that are brought up by operations researchers to justify the neglect of the psychological cost of waiting. One is that psychological considerations are too fuzzy to be captured in analytical models. Another is that even if such models could be constructed, their prescriptions would not differ qualitatively from those of traditional queuing research. In this paper we dispute both arguments. We illustrate that some aspects of the psychological cost of waiting can be incorporated into an analytical queuing model. We also show that analysis of such a model can result in prescriptions that are inconsistent with those dictated by conventional queuing analysis.

It is thus suggested that service should be scheduled in a manner that not only considers the time spent in the system, but also considers consumers’ (dis)satisfaction (for a review of research on consumer satisfaction, its measurement and validity, see Yi 1990). The latter criterion can, for example, help explain why it may be advisable for popular restaurants to hand out menus and ask for their patrons’ orders as they wait to be seated, thus providing a portion of the service at an early stage of the wait. Note that this practice could also reduce the waiting time and generate additional revenues, but even if this were not the case, in some contexts it would still be worthwhile. Similarly, when a patient awaits a doctor’s appointment, having a nurse or a receptionist perform part of the service, such as writing down the patient’s name and the purpose of the visit, may serve to lessen the patient’s dissatisfaction with waiting, even if the reduction in the sojourn time is
negligible. Even providing a service that is not actually needed may prove helpful, as long as it is not perceived merely as a manipulation attempt.

Extensions of the specific analysis we presented in this paper could examine additional functional forms of the service and the waiting functions. For example, in some situations, the service function may also depend on the timing and the duration of previous service portions. In such situations, the optimal policy is likely to involve "spreading" the service across a substantial portion of the waiting period. On the other hand, dividing service into minuscule segments can be a nuisance to customers, for example. Another extension could be to incorporate the effect of consumers' expectations regarding the remaining waiting duration into the cost function. Yet another important issue would be to focus on minimizing the maximal momentary evaluation in the course of the service (which affects the likelihood of balking) rather than the final dissatisfaction level. An analysis that would combine the two criteria (i.e., the global and the momentary), and optimize both concurrently, would be important as well. A general analysis of situations in which the sojourn time cannot be assumed to be constant is also needed.

More generally, the major purpose of this paper has been to draw attention to research about consumer service queues. We believe that the importance of an approach that accounts for psychological considerations cannot be overstated. Unfortunately, much of the voluminous research on queuing to date has not considered this a high priority. While it is clear that additional behavioral queuing research is required, a substantial base of knowledge has been accumulated across various research disciplines. Particularly relevant research has been conducted in the fields of psychology, marketing, economics, and sociology. Although there is no comprehensive literature review that is directly relevant, portions of the following may prove helpful: Fraisse (1963, 1984), Schwartz (1975), Jacoby et al. (1976), Levin and Zakay (1989), Hall (1991), Carmon (1991), and Carmon and Kahneman (1993).

We find it odd that while comparable negligence of psychological reality in the field of economics has been widely criticized (e.g., see the proceedings of the "The Contrast Between Psychology and Economics" conference, Hogarth and Reder 1986, and Thaler 1991), this has not been the case for the field of operations research (see Rothkopf and Rech 1987, Larson 1987, Green and Kolesar 1987, and Hall 1991 for rare exceptions). Failure to consider determinants of consumers' perceptions of queuing systems may at best result in inefficient systems, and in some cases lead to solutions of queuing related problems that simply do not work (see Larson 1987, for several enlightening examples of such embarrassments to conventional Queuing Theory).

Overall, psychologically based queuing research represents an exciting opportunity for researchers in a variety of fields. It is a domain in which surprisingly little research has been conducted, though we believe that this is bound to change soon.\footnote{J. George Shanthikumar is supported in part by the NSF Grant DDM-9113008 and by a grant from the Committee on Research, University of California, Berkeley.}

Appendix

\textbf{Lemma A1}. Let $\phi_i$ be the solution to $(d/dx)\phi_i(x) = r_i(\phi_i(x))$ with $\phi_i(0) \geq \phi_2(0)$ for $i = 1, 2$. If $r_1(s) \geq r_2(s)$, then $\phi_1(x) \geq \phi_2(x)$, $x \geq 0$.

\textbf{Proof}. We will prove the above theorem for the case $r_1(s) > r_2(s)$. The case $r_1(s) \geq r_2(s)$ can be then proved by a limiting argument. The proof is by contradiction. Suppose $\hat{x} = \inf \{ x : \phi_1(x) < \phi_2(x), x \geq 0 \}$ is finite. Then by the continuity of $\phi_1$ and $\phi_2$ one sees that $\phi_1(\hat{x}) = \phi_2(\hat{x})$ and that there exists a decreasing sequence $x_1, x_2, \ldots$, such that $\phi_1(x_n) < \phi(x_n)$ and $x_n \to \hat{x}$ as $n \to \infty$. Therefore

$$\frac{\phi_1(x_n) - \phi_1(\hat{x})}{x_n - \hat{x}} < \frac{\phi_2(x_n) - \phi_2(\hat{x})}{x_n - \hat{x}}.$$\hspace{1cm} (1)

Taking the limit of the above inequality as $n \to \infty$, one gets $r_1(\phi_1(\hat{x})) \leq r_2(\phi_2(\hat{x}))$. This contradicts our assumption thus proving the lemma.

\textbf{References}


Accepted by Linda Green; received September 1992. This paper has been with the authors 4 months for 1 revision.