OR FORUM

PERSPECTIVES ON QUEUES: SOCIAL JUSTICE AND THE PSYCHOLOGY OF QUEUEING

RICHARD C. LARSON

Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received January 1987; revision received April 1987; accepted August 1987)

PERSPECTIVES ON QUEUES: SOCIAL JUSTICE AND THE PSYCHOLOGY OF QUEUEING

RICHARD C. LARSON

Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received January 1987; revision received April 1987; accepted August 1987)

Queues involve waiting, to be sure, but one's attitudes toward queues may be influenced more strongly by other factors. For instance, customers may become infuriated if they experience *social injustice*, defined as violation of first in, first out. *Queueing environment* and *feedback regarding the likely magnitude of the delay* can also influence customer attitudes and ultimately, in many instances, a firm's market share. Even if we focus on the wait itself, the "outcome" of the queueing experience may vary nonlinearly with the delay, thus reducing the importance of average time in queue, the traditional measure of queueing performance. This speculative paper uses personal experiences, published and unpublished cases, and occasionally "the literature" to begin to organize our thoughts on the important attributes of queueing. To flesh out more of these issues, the author asks for your cards and letters.

"... a day full of waiting, of unsatisfied desire for change, will seem a small eternity."

—William James, 1891

We start this story at a local department store. After purchasing a red bike for my older son, Erik, I was given the sales slip and told to proceed to the inventory/checkout window, to give a copy of the slip to a clerk behind the window, and to await my son's new bike. Upon arriving at the window, I noticed a woman who was on the verge of tears. I questioned her and discovered that she had been waiting there 30 minutes for her merchandise, during which time numerous other customers had come and gone, carrying away with them their purchased items. Soon I learned first hand of the travails of my beleaguered friend, as I watched numerous customers who arrived after me successfully pick up their waffle irons, quilts, or automatic coffee makers and leave the store, typically within several minutes after arriving at the checkout window. Approximately 35 minutes later I was given a box containing the red bike, after which I left, with my frustrated friend still anguishing over her ever-increasing delay. I was so mad when I got to my car that I promised my family I would never open that box, but rather would return it unopened the following Saturday and purchase a different bike at a respectable bicycle shop, one where I could get good personal service and a higher quality product. I had gone to this department store in the first place for ease of selection and rapid service, objectives that clearly were not met. The following Saturday, I carried out my promise.

Social Justice

At the time of the bike experience, several students and I had just started some research on queues as perceived and experienced by customers. Our research was based on a single conjecture:

For the great majority of queueing system "customers" the actual and/or perceived utility of participating in the system is (1) a nonlinear function of queueing delay, and (2) multiattributed.

The bike experience added an important attribute: *social justice*, as measured by adherence to (or violation of) *f*irst *i*n, *f*irst *o*ut (FIFO). Other personal experiences and documented cases led to other attributes.

Subject classification: 531 the many attributes of queues, 572 queueing in service industries, 681 queues as experienced by customers.

0030-364X/87/3506-0895 $01.25

The following discussion is intended to share some preliminary thoughts about queues as experienced by customers and to ask for your help. The "literature" on this subject is scattered and not well organized. Any suggestions regarding literature not cited or case studies that reveal new insights would be most appreciated. Please send your comments to the author.

Slips and Skips. As a consequence of the bike experience, we defined in our ongoing research "slips and skips," two quantities whose magnitudes measure social injustice in queues. Imagine you join a queue at time zero. Another individual joins the queue some time later, but enters service before you. You have been victimized by a *slip*, as the second arriving customer has slipped by you. From the point of view of the second arriving customer, he has *skipped* over you. He who experiences a slip is victimized; he who skips gets a certain sense of satisfaction from his good fortune. If we consider an m-order skip (i.e., one created by a customer who skips over m customers) to be m skips, then for every slip there is a skip, and we have immediately a "theoretical" result that the total number of slips equals the total number of skips in any queueing system.

Slips and skips can be measured in different ways, such as queue slips, service slips, or system slips, depending on whether the injustice occurred in the queue, in service or within the entire system comprising both queue and service. If B skips over A in queue but A leaves service before B, then because of B, A has experienced a slip in queue, a skip in service, and neither a slip nor a skip for the entire system. Queueing theorists and social scientists have long believed that first come, first served (FCFS) is the socially just *queue* discipline and first in, first out (FIFO) the socially just *system* discipline.

An MIT doctoral student, Ethel-Sherry Gordon, and I derived the probability laws for slips and skips in a number of different popular queueing systems. These include parallel operating $M/M/1$ and $M/M/k$ systems and the $M/G/\infty$ system, which we see as a good model for the merchandise pickup window described previously. (From our observations, each sales slip presented at the window was immediately given to one of "many" storeroom clerks, a busy clerk having responsibility for locating merchandise for only one customer at a time, thus motivating the "infinite number of servers" approximation.) The technical work is described in Gordon and Larson (1987). In additional related technical work, Ward Whitt (1984) argues for more intensive analysis of slips and skips, which he calls "overtaking," within the context of data communications networks.

Social Injustice in Practice. In customers' perceptions of queues, fear of social injustice can often dominate queue waiting times. For instance, an accomplished management science consultant to the fast food industry has reported that customer satisfaction in certain single-queue, Wendy's restaurants is higher than in many multi-queue Burger King and McDonald's restaurants averaging half the queue waiting time as Wendy's. He believes the Wendy's customers prefer the longer queue with guaranteed first-come, first-served queue discipline to an "undisciplined" multi-line situation with high chance of social injustice (Lewin 1986).

Sometimes efforts directed at reducing queue delay may exacerbate social injustice. An example is my hometown supermarket, which opens additional cash registers whenever the checkout lines "become too long." The problem is that I always seem to be the one near the head of the line, with the most time invested in the queue; the "newcomers" behind me scurry over to the new register, entering service approximately in a last-come, first-served manner. Infuriating!

The threat of slips can have significant dollar consequences. One example is found with barge traffic on inland U.S. waterways. As tows proceed from one lock to the next, for instance on the Ohio and Mississippi Rivers, it has been a common practice of tug captains to proceed at high, fuel-inefficient speeds. Each captain races his tug to the next lock in an attempt to minimize the possibility that a tow behind him will overtake him en route and thus enter the next queue before he does. Incurring such a slip would delay the departure time at the next lock, thus lengthen the entire voyage time, thus cost extra dollars. Delays at congested locks can range from a few hours to more than 1 day. Fuel consumption of a tug is approximately proportional to travel distance times the square of the speed. A modest speed reduction from, say, 6 mph to 5 mph could save 31% in fuel consumption. The "anti-slip" proposal, put forward by the consulting firm of Ketron, Inc. (Kettelle 1986), was to preassign queue positions to tugs. Whenever a particular lock is "congested" (for example, when there are delays of 6 hours or more), a tug leaving an adjacent lock for the congested lock would at the moment of departure be assigned its position in queue at the congested lock. This practice would eliminate the threat of slips and skips, and thus the motivation for speeding. Ketron estimated that for a typical congested

lock the potential fuel savings would exceed $1,000,000 per year (Ketelle).

Slips and skips can also have disastrous effects on assembly lines. In a typical automobile line, cars are sequenced by some measure of similarity, such as "the next eight all get air conditioning" or "the next ten get the 'super option' package." The problem arises when the cars reach certain parallel service channels such as paint shops and exit from parallel service "out of sequence." Then, for instance, the "eight getting air conditioning" may be interspersed with several not getting air conditioning. The result significantly increases "set-up" and "set-down" costs.

My favorite case history whose outcome could be explained in terms of slips and skips involves an airline serving an airport in Texas. Passengers disembarking from eight or so flights that arrived in Houston between 7:00 and 9:00 A.M. complained loudly and vehemently about lengthy luggage handling delays. The vice president in charge of operations conducted several studies, employed consultants knowledgeable in queueing theory, and even hired additional baggage handlers, so that the total baggage delay never exceeded 8 minutes (an accepted industry standard) and yet the passenger complaints continued unabated. A closer analysis of the problem, which required simultaneous on-site observation by several researchers, revealed that the waiting time until luggage delivery consisted of two components: a 1-minute walking time from the aircraft to the luggage carousel and a 7-minute waiting time at the carousel. Most individuals on this early morning flight were businessmen flying in to get a head start on the business day in Houston. As passengers disembarked from the aircraft and approached the carousel area, a certain fraction of them (those with hand luggage) proceeded directly to the taxi stand, boarded a taxi, and commenced their working day; those waiting at the carousel were afforded the opportunity for seven minutes of watching passengers who disembarked after them start their business day before them. The customers' aggravation could be explained largely in terms of slips and skips. Those who were victimized by slips complained; those who enjoyed skips said nothing. The solution to this problem was to deliberately reinsert delays in the system. The aircraft disembarking location was moved outward from the main terminal, and the most distant carousel was selected for delivery of luggage, so the total walk time was increased from one to six minutes. After this insertion of delay was successfully completed and the system was perceived to be more socially just, passenger complaints dropped to nearly zero. We see here an example in which social injustice clearly dominates time in the system, the single measure often used by queueing traditionalists. Martin (1983), who reported this case, calls the solution an example of "perception management."

Environment

Banks provide a "textbook" setting for queues. A bank near my office in Cambridge, Massachusetts, in advertising for new tellers, implores prospective candidates to help shorten customer waiting times (Figure 1).

But waiting time (or line) reduction may not be as important as imaginative lobby design options. The Manhattan Savings Bank is apparently one of the most successful and rapidly growing savings banks in the City of New York. Their customer happiness does not depend on an above-average number of tellers, or on new computer technology (such as automatic teller machines), but rather on the fact that every business day from 10 A.M. to 2 P.M., in most of the bank's 16 New York offices, the bank offers live entertainment. Most often the entertainment is in the form of music provided by pianists and organists. However, the bank has now instituted such annual lobby-centered events as week-long, pure-bred dog exhibits, cat shows and a Christmas ice show. Customers no longer dread going to the bank and waiting in line to execute their financial transactions: they view the time they spend in the lobby as a positive and usually entertaining experience, so much so that on at least one occasion an enterprising entrepreneur sold admission tickets to the bank (unbeknownst to the bank) (Miller 1984).

Eliminating Empty Time. Unless engineered otherwise, waiting in queue can be a very negative and frustrating experience, even in the absence of social injustice. As a *Time* essayist recently said,

> Waiting is a form of imprisonment. One is doing time—but why? One is being punished not for an offense of one's own but for the inefficiencies of those who impose the wait. Hence the peculiar rage that waits engender, the sense of injustice. Aside from boredom and physical discomfort, the subtler misery of waiting is the knowledge that one's most precious resource, time, a fraction of one's life, is being stolen away, irrecoverably lost.
>
> ... Waiting can seem an interval of non-being, the black space between events and the outcomes of desires. It makes time maddeningly elastic, it has a way of seeming to compact eternity into a few hours (Morrow 1984).

William James, in his classic essay on the perception of time, argues that filled time appears to pass more quickly than empty time.

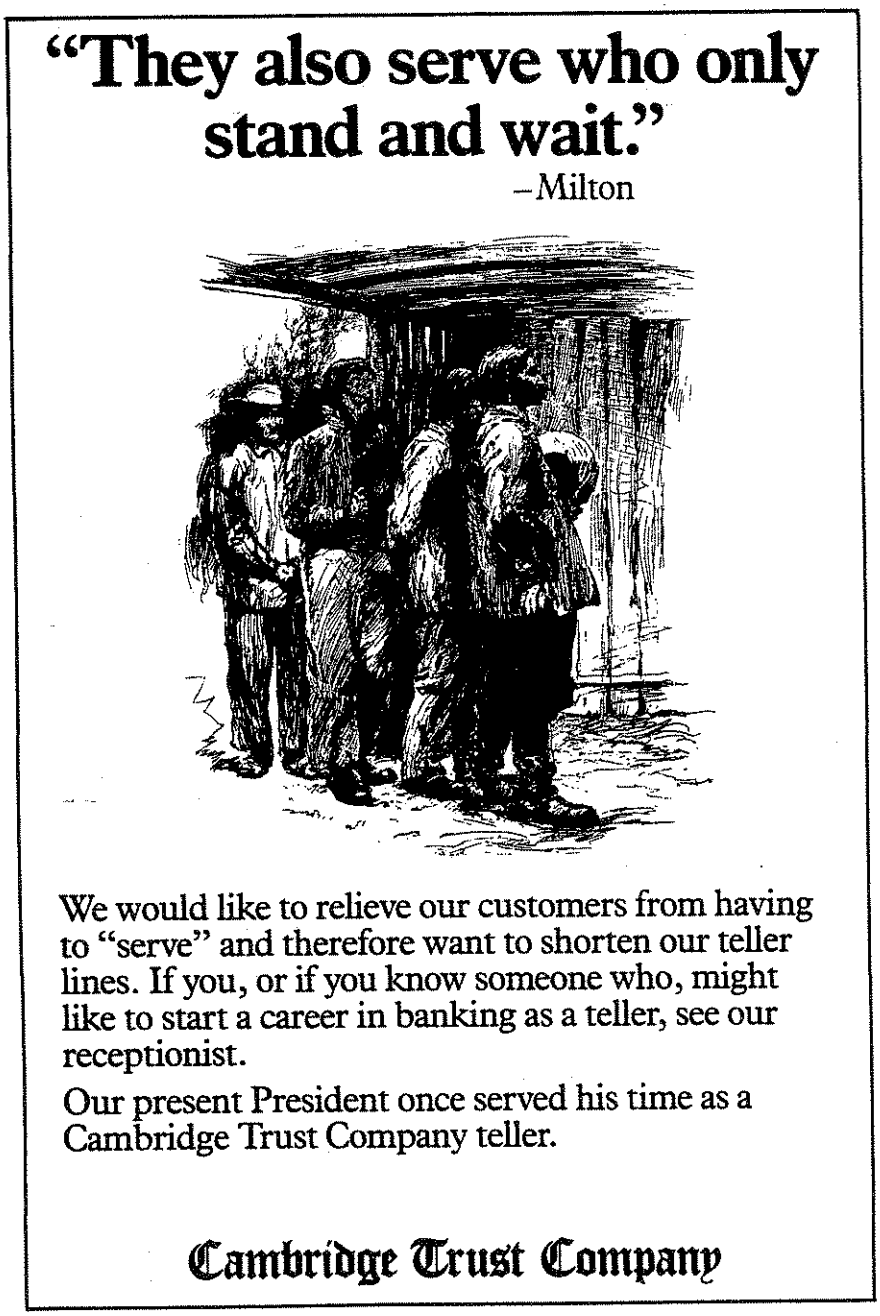

Figure 1. An advertisement by the Cambridge (Massachusetts) Trust Company for new tellers.

Taedium, ennui, Langweile, boredom, are words for which, probably, every language known to man has its equivalent. It comes about whenever, from the relative emptiness of content of a tract of time, we grow attentive to the passage of time itself,—expecting, and being ready for, a new impression to succeed; when it fails to come, we get an *empty time* instead of it, and such experiences, ceaselessly renewed, make us most formidably aware of the extent of the mere time itself (James 1891, p. 410; emphasis added).

The Manhattan Savings Bank represents an example in which a change in the queueing *environment* made the waiting experience a positive one. Customers standing in line had something to occupy their time. No longer were fractions of their lives perceived as being "wasted;" to the contrary, some were willing to pay for the entertainment.

Transportation planners have quantified the miseries of "empty time" waiting. They have found that bus passengers perceive a minute of delay at curbside waiting for a bus to "cost" two to three times that of a minute of time spent in the bus (Benakiva and Lerman 1985). This higher cost of waiting is used in models to design bus routes.

Another bank experience was reported by Martin. Shortly after a California bank installed computer terminals next to each teller (with the intention of

speeding service), numerous customers from at least one branch became so frustrated that they cancelled their accounts, and opened new accounts at a non-computerized bank "across the street" in which the teller service time averaged twice that of the first bank. In the computerized branch, most of the disgruntled customers were laborers who were depositing their Friday paychecks on their 12 noon to 1:00 P.M. lunch break. About 90 percent of the 30-second mean teller service time was spent in "wait-for-computer-to-respond" mode, caused by the lunch hour overload in the computer system. In the second bank, the tellers operating in a manual system were perceived as always busy during their average service time of 60 seconds; customers were happier in the second bank. Martin's solution was to change the queueing environment in the first bank:

> The clocks were replaced by green display terminals which gave the time, the weather forecast, the latest sports scores, publicity and interest rates on deposits. Waiting lines were combined into a single line feeding all tellers. Two TV monitors were installed conspicuously in the waiting area. Finally, a partition was erected between the counters and the terminals making the terminal invisible to the customers.
>
> The cashier was therefore always perceived to be busy. With the green displays and the TV monitors, the "demand" for customer time was high. Soon complaints went down significantly and the whole system became a sort of drawing card. The bank popularity went up significantly, a fact which certainly contributed to its increased profitability. As for the on-line banking system, it remained unchanged (Martin 1983).

One component of Martin's solution was to combine separate waiting lines into "a single line feeding all tellers." This approach eliminated the possibility of slips and skips in queue. Thus the "success" of the solution combined elements of both environment and social justice.

There are many other examples of waits that were perceived as empty or even aggravating that innovators transformed into positive experiences. A couple waiting out the last few hours prior to the birth of their child are not sharing pain and agony, they are jointly participating in Lamaze exercises. Most restaurants do not have queueing waiting rooms, they have cocktail lounges. Captive Audience TV (an Ohio-based corporation) attempts to entertain, as well as to market products by advertisements to, both adults and children standing in line in amusement parks. In a Mexican branch of the Republican National Bank of Texas, Walt Disney cartoons are used to entertain waiting customers.

Russel Ackoff, in the 1950s, emphasized the importance of elevator queue environment, an example that has become part of OR folklore. Floor-to-ceiling mirrors adjacent to elevators in high-rise hotels allow those who are waiting to fix their ties, comb their hair, and even perhaps coyly flirt via the mirror with others who are likewise waiting. According to Ackoff, those hotels that invested in such mirrors received far fewer complaints about elevator delays than competitors who did not (Ackoff 1987).

Selling to Captive Audiences. Entrepreneurs are beginning to recognize the potential for marketing goods and services to those standing in queue. In the Soviet Union, a study published by *Pravda* calculates that Soviet citizens waste 37 billion hours a year standing in line to buy food and other basic necessities (Morrow). In the United States, if we estimate that 200 million Americans occupy queues on an average of 30 minutes per day per person, we arrive at roughly 37 billion hours per year spent in standing in line in the United States. This figure is clearly speculative. However, to this author, who admittedly lives in a traffic-congested city, the figure of 30 minutes per day per person (when time spent at traffic lights, post office queues, in government bureaucratic offices, and so forth is added to the time spent waiting to purchase daily necessities) seems exceedingly conservative. Since by some estimates the average American watches approximately 4 to 5 hours of television per day, the time spent in queues would appear to be within an order of magnitude of the time spent watching television. The private sector spends approximately 25 billion dollars a year in television advertising, airing commercials which viewers may choose not to watch. It would seem that 2 to 3 billion dollars spent on marketing products to people in queues would not be inappropriate, considering that these individuals usually have very little in the way of alternatives to divert their attention.

Recognizing queues as captive audiences, we see in many cities in the United States and abroad various kinds of "street-level entrepreneurs" who earn money from motorists stopped in traffic queues. These include individuals selling flowers or newspapers (Boston), panhandlers who clean your windshield whether or not you want it cleaned (New York City), and street entertainers who perform such extravagant feats as breathing kerosene-fueled fire from their mouths (Mexico City).

The idea of changing empty time into useful time is of course the whole rationale behind marketing mobile cellular telephones, where businessmen and

women can carry on negotiations, make sales contacts, and perform other business activities while stuck in rush hour traffic.

Whatever the precise setting, it seems clear that the environment in which queue waiting occurs plays a fundamental role in a customer's perceived and/or actual cost of participating in that system. It seems, too, that a bit of ingenuity that would cost a minuscule fraction of the total operating cost of a facility often can go a long way toward alleviating customer anguish and discomfort, perhaps even transforming it into well being and happiness.

Feedback

From my observations, customers usually "feel better" about queueing when they are provided with information that allows them to estimate in advance their waiting time in queue. Individuals waiting in a lobby for one of several elevators can occupy their time watching the dials or lights moving as the respective elevators change floors throughout the building; sometimes one can even play a game with oneself, guessing which elevator will be the first to reach the lobby floor. A well-known international petroleum corporation directed some of its service station attendants to stand at the gasoline pumps with arm extended holding a pump's hose in order to indicate dramatically the total lack of queueing delay that would be experienced by customers entering the facility.

Disney World and Disneyland provide signs at points along the queueing channels to the various amusements indicating anticipated delays from those points. Such feedback helps customers choose which queue to enter, and it helps parents to select a strategy from child psychology for keeping their children "in line!" (For those who would like to learn more of Disney's "world class" management of queues, and how best, as a customer, to experience them, the author recommends Sehlinger and Finley 1985.)

For those of us who have been stuck in an aircraft on the ground waiting clearance for takeoff, I would conjecture that passengers experiencing a 30-minute wait without any feedback from the pilot are much more aggravated than those who are told at the beginning of the wait of the approximate 30-minute delay.

In the area of police response to calls for service, studies conducted in several cities—Worcester, Massachusetts; Wilmington, Delaware; and Kansas City, Missouri, to name a few—have surveyed citizens who have called 911, the police emergency response number. (See, for example, Cahn and Tien 1981 and McEwen, Connors and Cohen 1984). In attempting to "manage demands for police service," many police departments are now attempting to implement a "differential police response strategy" that deliberately delays certain lower priority calls for service by one-half to 2 hours (even in the presence of available servers, i.e., patrol cars) in order to leave servers available for potential near-term high priority incidents and to perform other important police duties. Extensive surveys of citizens who reported these lower priority incidents have shown that these "customers" are not dissatisfied with police service, even if delayed an hour or more, provided they are told of the estimated magnitude of the delay while on the phone, as well as the reasons for that delay. Thus, a citizen who waits, say, 60 minutes for the arrival of a police car, and who has been told, "Because of the current busyness level of the police force, Ma'am, we expect that a police car will be there in approximately one hour," is much more satisfied than a caller who is told, "A police car will be there right away, Ma'am," and who then ultimately waits 60 minutes. It's this latter customer who is likely to write irate letters to the editor of the local newspaper.

Motivated by the "police customers" attitude findings, in which deliberately inserted delay was acceptable, Christian Schaack and I undertook related modeling research on "cutoff priority" queues. These are queues in which preemption is not allowed and in which certain lower priority customers are deliberately delayed in queue even in the presence of (a "few") available servers, in order to preserve a "rapid service" capability for high priority customers who may arrive in the near-term. We are currently proposing one of our models as an analytical tool to help police planners design differential response strategies (Schaack and Larson 1986a, b).

Feedback need not always be provided directly by technology or system personnel. Sometimes it can be indirect. For instance, one might conjecture that a customer waiting in queue would have a "better experience" entering a queue behind 10 individuals, each of whom was observed to require precisely 1 minute of service time, rather than behind one individual who eventually required 10 minutes of additional service time. The conjecture is that the steady observed queueing delay "progress" experienced in the former case is in some sense psychologically more comforting than the uncertainty associated with not knowing when the single customer ahead would be completed. The movement of each of the 10 customers in the first line is providing feedback to our tagged customer that his likely total waiting time will be approximately 1 minute per customer ahead of

him. The second tagged customer has no such assurance, as there is no evidence of progress in this queue until the customer in service finally leaves.

To summarize our discussion to this point, we have attempted to argue or demonstrate that at least three attributes other than queueing delay play key roles in a customer's queueing experience: social justice, queueing environment, and feedback about delays. But what about the queueing delay itself? This is the subject of the next section.

Nonlinearity

In the 1960s, the Boston Police Department answered telephone calls for service as follows: each operator had before him (her) an identical toggle switchboard, with each toggle switch representing a potential incoming telephone call. Next to each switch was a small green lamp bulb. A blinking bulb signified that a given caller was in queue, waiting for his (her) phone call to be answered; a continuously illuminated bulb indicated that the respective caller was currently connected, speaking with one of the operators. During periods of congestion, particularly Friday and Saturday evenings, there were often more than 5 or even 10 "blinking green lights" at one time. Since operators could not be expected to recall the order in which the lights started blinking, they simply switched in at random one of the "blinking green lights" when they became available to handle another call. In effect, they were implementing a queue discipline of *s*ervice *i*n *r*andom *o*rder (SIRO).

Queueing Equivalent of EMV'er. Those who focus on mean time spent in queue (W_q) would not be concerned with Boston's SIRO queue. As is well known in queueing theory, a wide class of work-load-conserving systems enjoy the same mean queueing delay, independent of queueing discipline (e.g., FIFO, LIFO (last in, first out), SIRO, and so forth). However, as is also well known, the effectiveness of urban emergency response systems depends in a nonlinear way on system response time (for example, the time between calling the emergency response system and the arrival of appropriate service at the scene of the reported incident). In policing, for example, the probability of arresting a perpetrator near the scene of the crime is highest within 1 or 2 minutes after the report of the crime and drops roughly exponentially to a limiting positive value after approximately 10 minutes (Isaacs 1967). For many structural fires the dependence of dollar damage on response time of the fire apparatus follows an S-shaped curve through three distinct phases: incubation, escalation, and maturation; arrival of the fire apparatus within the gently sloping incubation period will keep the dollar damages to a minimal feasible amount (Halpern 1979). In emergency medical services the report of a person having suffered a myocardial infarction (i.e., heart attack) indicates that on-scene professional emergency medical services should start within 5 minutes after the infarction or the probability of death is almost 1.0. In considering the heart attack victim's personal "disutility" of a 5-minute response delay versus a 2.5-minute response delay, it seems clear that the 5-minute delay is "more than twice as bad" than the 2.5-minute delay.

Much of the analytical work on queues, and perhaps most OR textbooks (including my own!), focus on finding the fundamental quantities associated with Little's law, $L = \lambda W$. Assuming a steady-state queue, L is the time-average number of customers in the system, and W is the average total time spent in the system by a random customer. Analogously, the formula $L_q = \lambda W_q$ relates the time-average number L_q of customers in the queue and W_q is the expected time spent in queue by a random customer. In fact, W and W_q have become two of the most fundamental quantities describing a queueing system's behavior. From a utility theoretic sense, however, both imply a linear disutility for waiting time, analogous to "EMV'er's" in money-oriented utility theory. From our point of view, one should seek not W or W_q necessarily, but the expected disutility of experiencing (as a customer) the entire system or simply the queue. Only in the special case of linear disutility of delay are the two calculations equal, subject to a positive multiplicative constant. (See, for instance, Keeney and Raiffa 1976.)

All of the nonlinear production functions associated with urban emergency services point to the need for deviating from linear measures of queueing delays and system response times. With regard to the SIRO police queue discussed previously, while the mean queueing delay may not be affected by the SIRO strategy, it is well known that this service policy significantly increases the variance over what would be achieved with a "socially just" FCFS queue discipline. Intriguingly, it is possible—depending on the particular production function and other system characteristics—that SIRO or LIFO may in fact be preferred to FCFS for increasing the chance of saving the heart attack victim or arresting the crime perpetrator.

The productivity of time-shared computer system users has been found to vary in a markedly nonlinear way with the system response time. Thadhani, in 1981, reported that as computer system response time

is pushed into the sub-second range, user productivity as measured in the number of user interactions per hour increases dramatically. This production function, which is reproduced as Figure 2, shows a marked "elbow" at about 0.5–0.8 seconds. The data refer to users of an IBM system 370, model 168, multiprocessor system, supporting programmers involved in software development. Thadhani argues that a hardware system with its associated user-queueing software should nowadays be designed to minimize user response time rather than maximize processor utilization, as had been the practice through the 1970s. Recognizing the "elbow jump" in user productivity, sub-second systems can dramatically increase overall system productivity, especially now that the dollar costs of the users of a fully loaded time-sharing system far outweigh the hourly operating costs of the system hardware, software, and maintenance personnel (Minicucci 1982).

Assessing Preferences. Not all nonlinearities in queueing are due to measurable production functions. Some are caused simply by a customer's growing feeling of "annoyance" (Palm 1953). To investigate some of our conjectures regarding disutilities of queueing delay, we have conducted (and are continuing to conduct) interviews to assess people's disutility functions for waiting times in queues. The objectives of the interviews have been:

(1) To investigate the belief that the disutility of waiting is indeed a nonlinear function of the amount of time spent waiting in the queue.
(2) To determine which specific member, if any, of a parametric family of disutility functions satisfies the qualitative and quantitative assessments of the individual.
(3) To investigate the conjecture that queueing environment and/or waiting time information alters one's disutility function.
(4) To begin to understand risk aversion versus risk proneness in queueing situations.

In the interviews, we asked the subjects to indicate their preferences for a queueing system under different queueing environments. The subjects then gave certainty equivalents for a set of lotteries, after which we

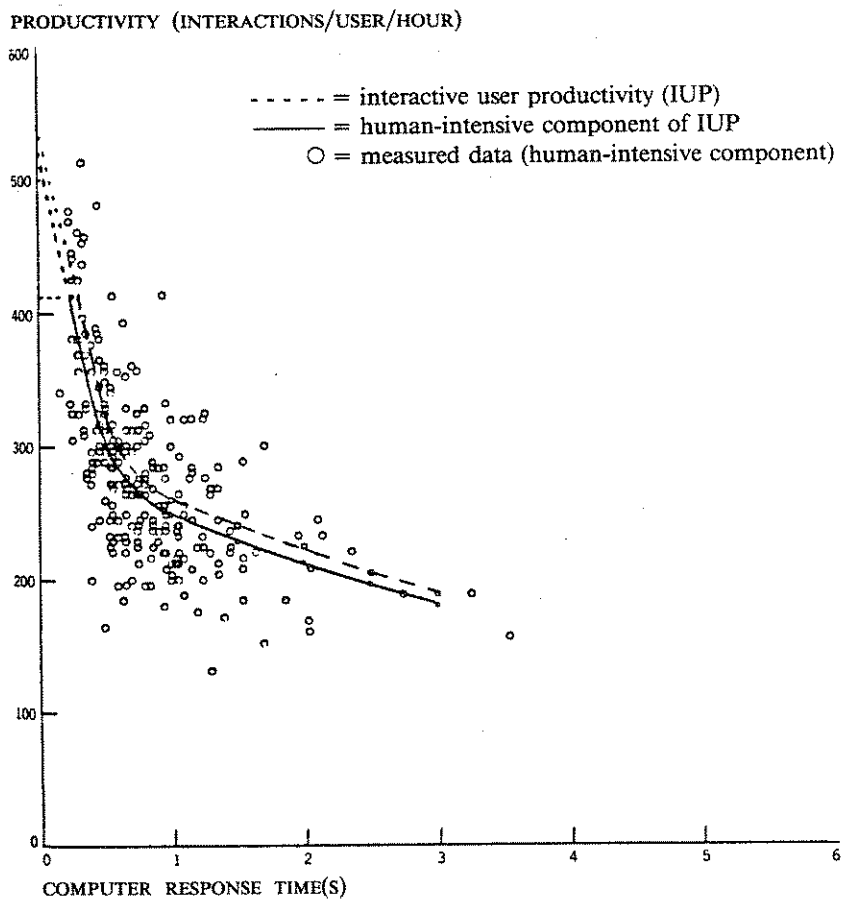

Figure 2. Interactive computer user productivity vs. computer response time (Thadhani 1981).

determined a least squares functional fit through those assessed points. Different sets of subjects were selected for each of the various queueing systems considered. The first scenario considered waiting for a bus, where the subject has an appointment scheduled at the end of the bus line. The second scenario involved purchasing food at a "fast food" outlet.

While the utility assessment details are somewhat standard and will not be included here for the sake of brevity, we state several illustrative findings:

- With one exception (in 10 interviews), all disutility functions varied nonlinearly with queueing delays.
- With one exception, each subject's disutility function depended on queueing environment.
- We found cases of risk proneness as well as risk aversion, sometimes in the same utility function.
- The disutility function seemed to depend heavily on the broader environment in which the individual was experiencing the delay (e.g., whether a hard time limit existed for the individual) (Leung 1984).

Clearly, these results are very preliminary. Additional interviews plus empirical observations need to be undertaken.

Conclusions

We believe that many attributes other than queueing delay contribute to a customer's utility or disutility in experiencing a queueing system. And even for the queueing delay itself, we believe that, for many if not most systems, the utility or disutility of experiencing the delay is a nonlinear function of the delay.

Our goal in this paper has not been to criticize those who advance the theory of queues. For instance, an amazing amount of work has addressed the need to "go beyond" the linear measures L and W. (As a partial list, see Whitt 1981, 1983a, b; Barnett 1978; Smith and Whitt 1981; Winston 1977; and Stidham 1970.) Nor have psychologists and sociologists been "idle," in queueing parlance. (See, for example, Fraisse 1963; Frankenhaeser 1954; Fraser 1966; Goldfarb and Goldstone 1963; Gulliksen 1950; Hirsh, Bilger and Heathridge 1950; Mann 1969; Orme 1969; Stroud 1955; and the summary in Bjorkman 1984.) Those of us who teach "queueing theory" in classrooms must remember to transcend Little's law and demonstrate the importance of higher moments and nonlinearities. But—with few exceptions—the attributes of queues other than wait (and related physical quantities) have been the subject of folklore, at best, not having benefited from systematic study. Some of these attributes may be perceived as "psychological" and in some sense too vague to be the subjects of careful analysis. I would suggest that the definition and analysis of "slips and skips" is a counterexample to this type of reasoning. Others may say that a customer's "attitudes" are extremely subjective and not nearly as important as rigorously measurable quantities. But as the marketing community has shown, attitude changes can cause customers to "brand switch," thereby substantially affecting corporate market shares. And queueing theorists, having a knowledge of customer attitudes, may find new ways to model queues and/or new queues to model (e.g., cutoff priority queues).

As we have attempted to show, customers' queueing experiences and attitudes can impact a wide range of firms, including fast foods, department stores, banks and hotels, transportation services, emergency services, "theme" parks, and airlines ("Up, Up, and Delay," a week-long special feature of ABC Television nightly news, April 13–17, 1987, covered some examples from the airline industry). Those in psychology and sociology have built up an impressive queue of results pertaining to attitudes toward waiting—the queue just waiting to be served by some OR people who wish to integrate mathematical methods with their empirical findings. Along these lines, I recommend a recent paper by Maister (1985), who offers 8 propositions regarding the psychology of queueing, each of which could be the focus of empirical verification and new mathematical modeling.

We have touched only the surface in this discussion. For instance, Arnold Barnett reports a type of worst delay "memory persistence" among passengers riding subways; that is, they perceive the service level to be near the worst experienced during the past week or month (Barnett and Saponaro 1985). Rothkopf and Rech (1987) argued recently that a widely advocated "queueing efficiency"—merging separate queues into a single combined queue—involves important issues beyond the standard reduction-of-mean-queue-delay-argument demonstrated by Erlang's formulas; in fact, if customers can know queue lengths prior to arrival and if they can jockey after arrival (presumably without "wreaking havoc" on the "social justice scale"), then many of the Erlang-derived advantages of combining queues apparently disappear and advantages of separate queues (e.g., express checkout lanes, servers' personal acquaintances with individual customers) may dominate. The 1973 U.S. gasoline crisis demonstrated that during goods shortages customers seem more attracted to longer queues than shorter ones, perhaps feeling that those in line have "inside

information" on impending stockouts. Hudson Hoagland (1966) showed empirically that one's perception of time passage varies with body temperature and conjectured that it also depends on a variety of other factors. Some "classes" of customers value their time more highly than others and are willing to pay to avoid or reduce queueing delays (see Kleinrock 1967 and Glazer and Hassin 1986).

Undoubtedly there are many factors, psychological, physiological and otherwise, that affect customers' perceptions of and experiences in queues. Here we have identified queueing environment, the level of information that one has about anticipated delays, and some measure of social injustice, recognizing there are other factors that remain undefined at this time. Better understanding of these relationships may have beneficial impacts on all relevant parties. Queue system managers may be able to find ways of reducing the disutility of queueing that are less expensive than the standard approach (which is to add more servers or to add technology to speed up service). Queue customers may have a more pleasant experience while in the queueing system. Firms seeking additional customers may be able to redesign their service facilities with an eye toward increased customer demand, in part through better understanding of how each prospective customer answers that proverbial question, "To Queue or Not To Queue?"

"I think the worst thing in the world is waiting," wrote "Thoughtful," in a "Confidential Chat" column in the *Boston Globe* (November 17, 1984). Of the six letter responses to "Thoughtful," one in particular is relevant to our discussion here, and I take the liberty of quoting it to conclude my remarks.

Dear Thoughtful:

I used to feel as you did about waiting. It was awful. I was so impatient. Now it is different because I am different. I use the time spent waiting to my advantage.

Here are a few of the things I do while waiting: I think about good things, projects I would like to do some time; I plan out the details in my mind. I pray instead of stewing because I have to wait. I read. (I usually keep a book or pamphlet with me.) I knit if it is going to be a long wait. I made seven afghans last year while I was waiting in hospitals. A side benefit was that I made a lot of nice acquaintances because people stopped to talk to me about what I was making.

To sum it up, I kind of make the time I wait work for me, and I keep it simple. A positive attitude and an openness to adventure also helps you expect something good to happen to you. You would be surprised at what you can see and learn and do while you wait!

Here's hoping you, too, can turn it around!

Queen of the Lilacs.

The author awaits your letters!

Acknowledgment

Our research on queues as experienced by customers is supported by the National Science Foundation, grant 8411871-SES. For their helpful suggestions, the author thanks A. Barnett, G. Bitran, D. Gross, K. R. Hammond, V. B. Iversen, J. Kettelle, A. Lewin, J. D. C. Little, O. B. G. Madsen, S. J. Pollock and W. Whitt.

References

ACKOFF, R. 1987. Personal communication to the author (April 3).

BARNETT, A. I. 1978. Control Strategies for Transport Systems with Nonlinear Waiting Costs. *Trans. Sci.* **12,** 119–136.

BARNETT, A., AND A. SAPONARO. 1985. Misapplications Reviews: The Parable of the Red Line. *Interfaces* **15,** 33–39.

BENAKIVA, M., AND S. LERMAN. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press, Cambridge, Mass.

BJORKMAN, M. 1984. Decision Making, Risk Taking and Psychological Time: Review of Empirical Findings and Psychological Theory. *Scand. J. Psychol.* **25,** 31–49.

Boston Globe. 1984. Confidential Chat: "Playing the Waiting Game" (Nov. 17).

CAHN, M. F., AND J. M. TIEN. 1981. An Alternative Approach to Police Response, Wilmington Management of Demand Program. U.S. Department of Justice, National Institute of Justice, Washington, D.C.

FRAISSE, P. 1963. *The Psychology of Time.* Harper & Row, New York.

FRANKENHAEUSER, M. 1954. *Estimation of Time.* Almquist & Wiksell, Stockholm.

FRASER, J. T. (ed.). 1966. *The Voices of Time, A Comparitive Study of Man's Views of Time as Experienced by the Sciences and by the Humanities.* George Brazillen, New York.

GLAZER, A., AND R. HASSIN. 1986. Stable Priority Purchasing in Queues. *Opns. Res. Lett.* **4,** 285–288.

GOLDFARB, J. L., AND S. GOLDSTONE. 1963. Time Judgement: A Comparison of Filled and Unfilled Durations. *Percept. Mot. Skills* **16,** 376.

GORDON, E. S., AND R. C. LARSON. 1988. Slips and Skips in Queues (in preparation).

GULLIKSEN, H. 1950. The Influence of Occupation upon the Perception of Time. *J. Exp. Psychol.* **69,** 561.

HALPERN, J. 1979. Fire Loss Reduction: Fire Detectors vs. Fire Stations. *Mgmt. Sci.* **25,** 1082–1092.

HIRSH, I. J., R. C. BILGER AND B. H. HEATHERAGE. 1950. The Effect of Auditory and Visual Background on Apparent Duration. *Am. J. Psychol.* **69,** 561.

HOAGLAND, H. 1966. Some Biochemical Considerations of Time. In *The Voices of Time,* J. T. Fraser (ed.). George Braziller, New York.

ISAACS, H. H. 1967. A Study of Communications, Crimes, and Arrests in a Metropolitan Police Department. Appendix B in Blumstein, A., et al., Task Force Report: Science and Technology, A Report to the President's Commission on Law Enforcement and Administration of Justice, U.S. Government Printing Office, Washington, D.C.

JAMES, W. 1952. *Principles of Psychology.* Holt, New York (originally published in 1891). (Opening quotation from Chap. XV, "The Perception of Time," 410.)

KEENEY, R. L., AND H. RAIFFA. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley & Sons, New York.

KETTELLE, J. D. 1986. A private communication of Ketron correspondence dated 1977.

KLEINROCK, L. 1967. Optimal Bribing for Queue Position. *Opns. Res.* **15,** 304–318.

LEUNG, P.-L. 1984. Multiattribute Queueing Theory. Master's thesis in Operations Research, Massachusetts Institute of Technology, Cambridge, Mass.

LEWIN, A. 1986. Private communication to author.

MANN, L. 1969. Queue Culture: The Waiting Line as a Social System. *Am. J. Sociol.* **75,** 340–354.

MAISTER, D. H. 1985. The Psychology of Waiting Lines, Chap. 8 in *The Service Encounter,* J. A. Czepiel, M. R. Solomon and D. F. Surprenant (eds.). D. C. Heath, Lexington, Mass.

MARTIN, A. 1983. Perception and Value Management. *Think Proactive* **8,** 95–101.

MCEWEN, J. T., E. F. CONNORS III AND M. I. COHEN. 1984. Evaluation of the Differential Police Response Field Test. Research Management Associates, Inc., Alexandria, Va.

MILLER, T. 1984. Eschewing Gifts, This Bank Offer Pianists, Dogs, Cats and Ice Shows. *Wall Street Journal* (May 6).

MINICUCCI, R. A. 1982. Sub-Second Response Time, A Way to Improve Interactive User Productivity. *SMC Newsletter 82–19* (November) pp. 1–10. Corporate Information Systems and Administration, White Plains, N.Y.

MORROW, L. 1984. Waiting as a Way of Life. *Time,* July 23, 1984, p. 65.

ORME, J. E. 1969. *Time, Experience and Behavior.* Cliffe Books Ltd., London.

PALM, C. 1953. Methods of Judging the Annoyance Caused by Congestion. *TELE* (English ed.), No. 2, 1–20.

ROTHKOPF, M. H., AND P. RECH. 1987. Perspectives on Queues: Combining Queues Is Not Always Beneficial. *Opns. Res.* **35,** 906–909.

SCHAACK, C., AND R. C. LARSON. 1986a. An *N*-Server Cutoff Priority Queue. *Opns. Res.* **34,** 257–266.

SCHAACK, C., AND R. C. LARSON. 1986b. An *N*-Server Cutoff Priority Queue Where Arriving Customers Request a Random Number of Servers. To appear in *Management Science.*

SEHLINGER, R., AND J. FINLEY. 1985. *The Unofficial Guide to Walt Disney World.* Menasha Ridge Press, Hillsborough, N.C.

SMITH, D. R., AND W. WHITT. 1981. Resource Sharing for Efficiency in Traffic Systems. *Bell Syst. Tech. J.* **60,** 39–55.

STIDHAM, S., JR. 1970. On the Optimality of Single-Server Queuing Systems. *Opns. Res.* **18,** 708–732.

STROUD, J. M. 1955. The Fine Structure of Psychological Time. In *Information Theory in Psychology,* H. Quastlen (ed.). Free Press, New York.

THADHANI, A. J. 1981. Interactive User Productivity. *IBM Syst. J.* **20,** 407–423.

WHITT, W. 1981. Comparing Counting Processes and Queues. *Adv. Appl. Prob.* **13,** 207–220.

WHITT, W. 1983a. Deciding Which Queue to Join: Some Counterexamples. *Opns. Res.* **34,** 55–62.

WHITT, W. 1983b. Untold Horrors of the Waiting Room: What the Equilibrium Distribution Will Never Tell About the Queue-Length Process. *Mgmt. Sci.* **29,** 395–408.

WHITT, W. 1984. The Amount of Overtaking in a Network of Queues. *Networks* **14,** 411–426.

WINSTON, W. L. 1977. Assignment of Customers to Servers in a Heterogeneous Queueing System with Switching. *Opns. Res.* **25,** 468–483.