There's More to a Line than Its WAIT

BY RICHARD C. LARSON

WHEN I bought a red bike for my son Erik, I received a sales slip and was told to give a copy of it to a clerk at the inventory/checkout window. Arriving at the window, I noticed a woman on the verge of tears. She had been waiting 30 minutes for her purchase while many other customers had come and gone. Soon I, too, saw many customers arrive and collect waffle irons, quilts, and automatic coffee-makers. Some 35 minutes later, I was given a box containing the red bike, and I left with my frustrated friend still anguishing over her ever-increasing delay. I was so mad that I returned the box unopened the following Saturday and purchased a different bike at a respectable bicycle shop with good personal service and a higher-quality product.

My shopping experience coincided with research on the way customers experience lines or "queues," work I'd just begun with colleagues at M.I.T. As an electrical engineer in this venture, I'm part of a tradition begun by Danish telephone engineer A.K. Erlang, who in 1917 invented mathematical queuing theory to help "size" telephone switching systems—that is, determine their capacity so the chances of getting a busy signal can be kept to an acceptable minimum.

Over the years, telecommunications has continued to be a primary application area for queuing theorists, with many productively employed at such world-class institutions as AT&T Bell Laboratories. Today's queuing engineers help design digital communications systems, schedule operators, and undertake related tasks, all with an eye toward reducing delays. Ever since its birth, queuing theory has also served broader service industries trying to decrease their customers' discontent. Banks try to devise efficient queuing systems to reduce customers' waiting time at the teller's window. Airlines try to get baggage to travelers as fast as possible. Police departments try to reduce the response time to emergency calls.

Much of this work has placed an undue emphasis on the average or mean delay in queue. Recently it has become clear that other factors may be more significant than mean waiting time. Social justice is one such factor. Are you getting served last even though you arrived first, as I and my friend were at the department-store window? Queuing theory has designated first in, first out as the measure of how socially just any particular queuing system is.

The waiting environment is a second factor. Are you sitting in traffic staring at a stalled car,

ILLUSTRATIONS: SUSAN LE VAN

"Queuing theorists" are starting to realize that what happens to you while you're in line is more important than how long you're there.

or are you in animated conversation with a friend? A third factor is what feedback you are given about delays. Does the airline say your flight will be an hour late so you can do something useful at home, or does it say the flight will be on time so you wind up wasting that hour at the airport? These apparently subjective considerations may turn out to be far more important than the older mean-time measure, not only in people's own feelings about waiting but in the overall performance of an entire queuing system.

Older theories also make the mistaken assumption that individual experiences—one person waiting 30 minutes and another 10 minutes, for instance—can be lumped to produce an average cost (in this case 20 minutes), since the idea is to reduce the mean wait. But the cost of a given wait to an individual may not be at all proportional to the time waited. The extreme case is an emergency medical response to a heart attack. Often treatment within two or three minutes will save the victim, but if treatment doesn't start within five minutes, death is almost certain. Clearly a five-minute delay is more than twice as bad as a two-and-a-half-minute delay.

What really matters is the cost of one's waiting experience, not just in money but in frustration, anger, and other stresses. If they understood this principle, industries like fast-food chains, banks, and airlines could reduce their customers' anxieties, better manage their own budgets, and even save lives.

Fairer Is Better

For customers, perceived social injustice can dominate the waiting experience. Arie Lewin, now at the National Science Foundation and formerly a management consultant to the fast-food industry, reports 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 waiting time. He believes the Wendy's customers prefer the longer queue with guaranteed first-come, first-served discipline to an "undisciplined" multi-line situation with high chances of social injustice.

Sometimes efforts directed at reducing queue delay may result in more discontented customers—and therefore poor queuing-system performance—by exacerbating social injustice. My supermarket opens additional cash registers whenever the checkout lines get too long. But I always seem to be near the head of the line when the "newcomers" scurry over to the extra register and pass through the checkout in a last-come, first-served manner.

These examples illustrate "slips and skips," magnitudes of which can be measured to yield an objective estimate of social injustice. You've been victimized by a slip when another person joins a queue after you but gets served before you, since that person has slipped by you. For every slip there is a skip (from the other person's perspective, you have been skipped over).

There are queue slips, service slips, or system slips, depending whether the injustice occurs in the queue, in service, or within the entire system comprising both queue and service. If B skips over A in a queue but A leaves service before B, then A has experienced a slip in the queue, a skip in service, and neither a slip nor a skip for the entire system. A queue that is first-come, first-served does not allow queue slips or skips. A system that is first in, first out does not allow system slips or skips. Queuing theorists and social scientists have long believed that first come, first served is the socially just queue discipline and first-in, first-out the socially just system discipline.

Threatened slips can have significant dollar consequences. Take, for instance, barge traffic on inland waterways. As tugs go from one lock to the next on the Ohio and Mississippi rivers, captains often proceed at high, fuel-inefficient speeds to minimize the possibility that those behind will enter the next queue first. Such a slip could delay the departure time at the next lock. Queue delays at locks can range from a few hours to over a day, and captains wish to avoid anything that could lengthen a voyage and increase its cost. Modestly slowing down, say from six mph to five mph, could save 31 percent in fuel consumption.

Ketron Corp., a Washington, D.C., operations research and consulting firm, made an anti-slip proposal that assigns queue positions to tugs. Whenever a particular lock is congested with delays of six hours or more, each tug headed there is assigned a queue

---

RICHARD C. LARSON, professor of electrical engineering and computer science at M.I.T., has studied lines at automatic teller machines and discount stores, as well as the queuing systems for police and ambulance dispatch. He has plenty of time to reflect on these matters as he drives ever so slowly through Boston's Sumner Tunnel at the height of rush hour. He would welcome feedback from readers on their experiences waiting in line. A longer version of this article appeared in Operations Research.

Does the airline tell you your flight will be late so you can do something useful? Or do you wind up wasting time at the airport?

position at the moment it leaves the adjacent lock. Ketron estimates that this system could reduce annual fuel costs by $1 million per lock in the system.

My favorite slip-skip case history involves an airline serving the Houston airport. Passengers disembarking from flights that arrived between 7:00 and 9:00 A.M. complained loudly and vehemently about long luggage-handling delays. The vice-president in charge of operations conducted several studies, employed consultants knowledgeable in queuing theory, and even hired additional baggage handlers so that the total baggage delay—the time between leaving the plane and picking up baggage—never exceeded the accepted industry standard of eight minutes. But the passenger complaints continued.

On-site observations showed that the waiting time for luggage delivery consisted of two components: a one-minute walk from the aircraft to the luggage carousel and a seven-minute wait at the carousel. Most individuals on this early-morning flight were trying to get a head start on the Houston business day, and anyone with hand luggage proceeded past the carousel directly to the taxi stand. Passengers at the baggage carousel had to spend seven minutes watching others who had disembarked later start their business day first. Those who were victimized by perceived slips complained. Those who enjoyed their skips said nothing.

The solution: a sleight of hand by which the delay was actually increased, while passengers' perceptions of it were transformed. The disembarking location was moved from the main terminal and the most distant carousel selected for luggage delivery, so that total walk time was increased from one to six minutes. After this delay was added, the system was perceived as more just, since people no longer had to watch others get ahead of them. As a result, passenger complaints dropped to almost zero. Perceptions of social injustice clearly mattered more here than the actual time passengers spent in the system.

Organists and Cat Shows

"*Taedium, ennui . . . boredom,*" wrote William James in "The Perception of Time," "are words for which . . . every language known to man has its equivalent. It comes about whenever, from the relative emptiness of content in a tract of time, we grow attentive to the passage of time itself."

As early as the 1950s Russel Ackoff, professor of systems science at the University of Pennsylvania, made the elevator environment part of queuing-theory folklore. According to Ackoff, high-rise hotels investing in floor-to-ceiling mirrors next to elevators allow the people waiting to fix their ties, comb

When the walking time to the luggage carousel increased and waiting time decreased, complaints dropped almost to zero.

their hair, and even flirt coyly with fellow time-servers. The queue-wise hotels in question, he wrote, received far fewer complaints about elevator delays than did their mirrorless competitors.

Alain Martin, a Toronto-based consultant specializing in perception and time management, reported that after a certain California bank tried to improve efficiency by installing computer terminals next to each teller, many customers canceled their accounts and opened new ones at a nearby non-computerized bank with twice the average service time. Most of the customers turned out to be laborers depositing their Friday paychecks on their noon lunch hour. The mean service time at the computerized bank was a scant 30 seconds, but the tellers appeared to be inefficient because they had to spend 90 percent of those seconds waiting for the computers, overloaded in the lunch-hour rush, to respond. At the second bank the wait was 60 seconds, but because the tellers seemed continuously busy, customers were happier.

Martin's solution: the computerized bank replaced clocks, which conveyed only the tedium of time's passage, with lively green display terminals showing time, weather forecast, publicity, bank interest rates, and the latest sports scores. The bank also added two TV screens in the waiting area and erected partitions to hide each terminal from customers so that tellers always seemed busy. In addition, the separate queues were made into one line feeding all the tellers—a simple way of avoiding slips and skips. Martin, who had worked on the Houston airport problem, called this a perception management situation, and solved it by combining environmental and social-justice principles.

Some banks have done even more spirited environmental end-runs around potential customer frustration. The happiness of customers at the Manhattan Savings Bank, one of the fastest-growing savings banks in New York City, depends not on extra tellers or new computer technology, but rather on the fact that there is live entertainment every day from ten to two in most of the bank's offices. To its original entertainment—piano and organ music—the bank has now added week-long purebred-dog exhibits, cat shows, and a Christmas ice show. So successful have these ventures been that, according to the *Wall Street Journal,* an enterprising individual has sold tickets to one of the shows—unbeknownst to the bank.

Entrepreneurs would do well to recognize the po-

tential for marketing goods and services to those standing in line. In the United States, if 200 million individuals spend an average of 30 minutes per day waiting in lines, that adds up to roughly 37 billion hours a year. (This figure is clearly speculative. But though I admittedly live in a traffic-congested city, 30 minutes per person per day seems conservative. Consider the time spent sitting at traffic lights, purchasing necessities, and waiting in post office lines and bureaucratic offices.) Since it is said that the average American watches four or five hours of television a day, the time spent in lines would appear to be perhaps a tenth as much. The private sector spends around $25 billion a year on TV advertising that viewers may choose to ignore, so $2 billion to $3 billion doesn't seem a high price to pay for marketing products to queue-waiters with no relief for their boredom.

The idea of changing empty into useful time is of course the rationale behind mobile cellular phones, by means of which business people carry out telemarketing and other activities while they're stuck in rush-hour traffic. Others stuck in traffic are using tape cassettes to learn foreign languages or listen to novels.

"I think the worst thing in the world is waiting," wrote "Thoughtful" in a "Confidential Chat" column on November 17, 1984, in the *Boston Globe*. Among the responses to Thoughtful's letter was the following: "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 . . . I plan out the details in my mind. I pray instead of stewing . . . I read . . . I knit . . . 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 . . . Here's hoping you, too, can turn it around!"

Signed, *Queen of the Lilacs*.

Knowledge Is Comfort

Knowing how long one has to wait doesn't change the delay, but it can certainly help relieve stress. Disney World and Disneyland post signs along

"I think about good things, I pray, I read, I knit. I kind of make the time I wait work for me."

On any given day subway riders perceive service to be near the worst experienced during the preceding week or month.

queuing channels indicating anticipated delays to the various amusements. At a conference I attended, a petroleum corporation was said to have directed some of its service-station attendants to stand at the pumps holding the hoses so customers would know they wouldn't have to wait to get gas.

Air passengers experiencing a 30-minute wait with no feedback from the pilot usually seem much more annoyed than those told at the beginning that they will have to wait a half hour. However, there is also the irritation of being told about a 30-minute wait when the actual delay is twice as long. (It would seem better for airlines to slightly overestimate delays: passengers would be pleasantly surprised at takeoff.)

Feedback doesn't need to be direct. A customer waiting in line might have a better experience entering it behind ten individuals, each of whom is observed to require precisely one minute of service time, than behind one individual who eventually requires ten minutes' additional service time. The hypothesis here is that the feedback of steady observed "progress" would convince customers they will enter service for sure after ten minutes of wait. This is clearly more comforting psychologically than not knowing when service will be initiated.

Responding to Emergencies

In any of the foregoing examples, it is irrelevant to use mean waiting time as a performance measure. The rules that determine who gets served next don't change the average wait: that stays the same no matter which system is chosen. Depending on the mean to measure response effectiveness in certain emergencies like crimes and fires isn't only insufficient, but potentially destructive as well.

The probability of arrest near the scene of the crime is highest within one to two minutes after the crime is reported, and it drops roughly exponentially until ten minutes have elapsed, at which point arrest probability levels off. For many fires in buildings, the dollar damage follows an S-shaped curve in which the two most important phases charted are incubation—the fire's slow beginning—and escalation—when the fire's rate of change, severity, and heat increases the fastest. If fire fighters arrive within the gently sloping incubation period, the dollar damages will be kept to a minimum.

In the 1960s the Boston Police Department got problematic results from a telephone answering system that relied on the mean-wait notion of queue performance. Each of up to 14 operators had to

work with an identical toggle switchboard on which each toggle represented a potential incoming telephone call. Next to every one of these switches was a small green bulb. A blinking green bulb signified that a caller was in line waiting to be answered. A continuously illuminated bulb showed that the caller was connected and speaking with an operator. During congested periods, especially on Friday and Saturday evenings, five to ten green lights would be blinking at the same time, and operators would switch from one to the next at random, since they couldn't recall the order in which the lights had begun to blink. The operators were implementing what queue theorists call service in random order.

What happened, of course, was that many people had to wait longer than others even though they had called in earlier. Clearly the random response system posed an unnecessarily high risk to the calling public. What was needed was greater social justice, but at the time there were no technological means for achieving this. Only in the late 1970s did digital technology make automatic call distributor systems possible. These systems, which can manage calls on a first-come, first-served basis, are now in use by the police departments in Boston, New York, and other major U.S. cities.

Police departments have also been paying attention to the idea of quick feedback I described above. In cities like Worcester, Mass.; Wilmington, Del.; and Kansas City, Mo., studies show that citizens calling 911 (the police emergency number) to report certain lower-priority incidents are rarely dissatisfied with police service if they are told the approximate magnitude of the delay they can expect and the reasons for it. Even delays of an hour or more appear to be acceptable. Many police departments are therefore trying to implement a "differential police response strategy" in which lower-priority calls are deliberately delayed a half-hour to two hours to leave patrol cars free for high-priority incidents.

Pondering the Imponderables

With few exceptions, queue characteristics beyond mean waiting time have been the subject of folklore and haven't been considered for systematic study. But while some researchers feel customer attitudes are subjective and therefore not rigorously measurable, it is also true, as marketing research shows, that attitude changes can make customers switch brands, thereby affecting corporate market shares. Subjective factors clearly can be measured with reference to such notions as slips and skips.

There are imponderables that go beyond what I've described here. For example, Arnold Barnett of M.I.T.'s Sloan School of Management reports a kind of worst-delay "memory persistence" among subway passengers. On any given day, they perceive the service level to be near the worst experienced during the week or month just preceding.

Michael Rothkopf, a senior staff analyst at the University of California's Lawrence Berkeley Laboratory, argues that merging separate queues into a single one—a strategy widely advocated for queuing efficiency—may be ineffective, since it so often depends on the standard reduction of mean delay. In fact, says Rothkopf, there are important issues that have little to do with the old standard measurement. If customers can know queue lengths before arriving, and if they can "jockey" for queue position after arriving without wreaking havoc on the social-justice scale, then separate queues like express checkout lanes may be the best solution. Or personal imponderables like the acquaintanceship of servers with individual customers may dominate. The 1973 gasoline crisis showed that during goods shortages, customers seem more drawn to long than short queues, perhaps because they feel those in line have inside information on impending stock-outs.

Understanding such subtle factors may do good all around. Queue-system managers may find less expensive ways to reduce queuing frustrations than the standard addition of servers or technology. Customers may have waits that are more pleasant. And firms looking for extra customers may redesign their services with an eye to better understanding how each of us answers the proverbial question, Is it worth the wait? □

RECOMMENDED FOR FURTHER READING

A. K. Erlang, "The Solution of Some Problems of Significance in Automatic Telephone Exchanges," *P.O. Electrical Engineering Journal*, p. 189, 1917.

Richard C. Larson, "Perspectives on Queues: Social Justice and the Psychology of Queuing," *Operations Research*, Nov./Dec. 1987.

Richard C. Larson and Amedeo R. Odoni, "Introduction to Queuing Theory and Its Applications," in *Urban Operations Research*, Prentice Hall, 1981.

D. H. Maister, "The Psychology of Waiting Lines," in *The Service Encounter*, D. C. Heath, 1985.

Alain Martin, "Perception and Value Management," in *Think Proactive*, PDI Press, Ottawa, 1983.

R. Sehlinger and J. Finley, *The Unofficial Guide to Walt Disney World*, Menasha Ridge Press, Hillsborough, N.C., 1985.