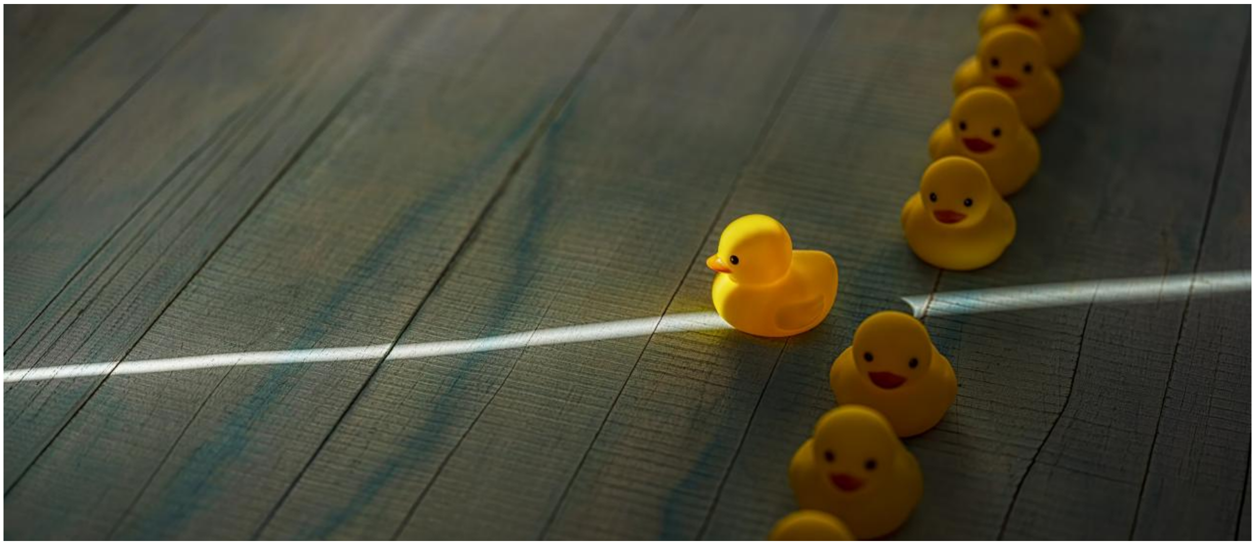
When Several Queues Are Better Than One



By Guillaume Roels , INSEAD Professor of Technology and Operations Management; Hummy Song, Assistant Professor of Operations, Information and Decisions, Wharton School; and Mor Armony, Professor of Technology, Operations and Statistics, Stern School of Business

One reliable method of queueing is less effective in knowledge industries. Here’s why.

We’ve all been in queues that seem to last forever, especially if we choose our queue at the checkout and the one next to ours is moving faster. You know the existential dread that comes along with standing in a dedicated queue and waiting interminably. To make service of all kinds more efficient, the predominant thinking in operations management is to form a single serpentine queue that feeds different servers – a pooled queue.

Traditional operations management theory has determined that pooling is more efficient. And it may be, if tasks or widgets are the items in the queue and it’s machines, not human beings, that are processing them. In a system with dedicated queues, it’s possible to have one that's empty and another queue that's full but no way to rebalance this. If the queue contains customers, naturally they can switch to the empty queue. But when we consider job assignments, for example, these can't just move across queues. So the dedicated queue is viewed as less efficient than a pooled one in terms of throughput and waiting time.

An impactful **paper** by Hummy Song and her co-authors focused on waiting rooms in emergency departments and found that when a part of the emergency department (ED) at a Kaiser Permanente hospital in California changed from a pooled queue to dedicated queues, patients had shorter wait times and a shorter length of stay. In the pooled setup, patients in the waiting room were assigned to a physician only when one became available. The switch to a dedicated system meant that as soon as patients were triaged, they were assigned to a particular physician and that physician's queue.

Interestingly the researchers found the opposite of traditional efficiency in queueing theory; patients had a shorter stay in the ED when they were in dedicated queues. Physicians anecdotally described how they felt more responsible in the dedicated setup for the people assigned to them in the waiting room before they actually saw them as a patient.

It's unusual in operations management to consider people in all their humanity, with their own idiosyncratic biases and behaviours. In "**Pooling Queues with Strategic Servers: The Effects of Customer Ownership**", forthcoming in *Operations Research*, we show that efficiency is improved across the system if organisations consider a concept that may be unfamiliar to scholars in this area: customer ownership. Service providers may develop a greater sense of obligation and accountability when they see all the customers in their queue as belonging to them rather than as an indiscriminate pool of demand.

We modelled this upending of queueing theory using customer ownership as the motivator. We described the split in servers' sense of customer ownership between when the customers enter the system and when they are right in front of the server. Our theory is human servers have human reactions that impact operational effectiveness – like how long someone spends in an ED.

When does a person become a customer?

When we talk about customer ownership, it's like a sense of responsibility that ED doctors had for people in the waiting room when they were triaged.

Other doctors may feel customer ownership when the patient is in front of them. In our model, we stripped out financial incentive notions – imagine call centre workers who get a bonus dependent on short wait times, for instance – to consider customer ownership on its own. (In fact, doctors at Kaiser are paid a fixed wage, so they have no financial incentive to see more patients.)

Organisational behaviour has documented a sense of **organisational ownership**, but customer ownership had not been previously analytically modelled nor had its consequences on process performance been considered.

In the model, we broke down customers who are already in the room versus the entire scope of the system. System-wide customer ownership is a combination of the people who are currently being served plus those still in the queue.

Servers either care about the customer they are currently serving or not only about that person, but future customers as well. Incorporated in customer ownership is an interesting time dimension, whether servers focus on the present or the future and how they behave.

The type of task matters

With a combination of game theory and queueing theory, one of the innovations of this paper is how we model the discretion that servers have in terms of their choice of the pace of work, which seems endogenous in practice.

In some cases, servers have very limited discretion. For instance, if you have to administer a survey of ten yes/no questions, you might have limited flexibility for taking much more or much less than the five minutes the survey was designed to last. But if the task is more knowledge-intensive, like physicians seeing a variety of cases in the ED, it's up to the server to decide how much time is needed. There is a clear distinction between the routine tasks where servers have some discretion and those that are typically more knowledge-intensive where servers may have more discretion about how much time they need to complete it effectively.

The type of service matters when choosing an efficient queueing system. With a standard type of task, the traditional theory that pooling queues are the most effective mechanism holds. But if the service provided is knowledge-intensive, it’s important to understand that the effect can be flipped.

We modelled the utility of servers and how their notion of customer ownership maximises it. This paper formalises what was observed in Song’s earlier work and demonstrates that the phenomenon can be justified on rational grounds. Our work is grounded in practice, and we built a theory to explain how it is transferrable to other contexts.

Our paper highlights the importance of accounting for human behaviour on the part of the server, shifting attention away from the customers and the human impact on process performance.

Broader implications of customer ownership

Queues aren’t only at the grocers or the airport. Managers in certain domains may need to consider redesigning their queueing systems not only when it comes to assigning customers to servers but also assigning work to team members. Another aspect to consider is the attention that individual contributors in knowledge-intensive services have on their own task queues. Think emails, assignments and other deliverables. Our paper suggests that in knowledge-intensive services where workers have a lot of discretion about the amount of time spent on a project, queues need to be managed a little bit differently. We find dedicating assignments to certain servers rather than pooling them to be more efficient.

Customer ownership is a concept that reflects organisational culture. As such, it can be modified, like other aspects of culture. Operations management often takes organisational culture for granted; our paper shows that operational design can shape it and thus impact performance. In particular, no one had previously pointed to queue configuration, which is an important operational lever, as a way to shape organisational culture. Yet switching to dedicated queues can lead to greater customer ownership.

When we think about queues, we usually think about them from the customer’s point of view. But we need to look at the human on the other end of the queue. Including a server’s customer ownership in consideration when planning queues will shorten the time for everyone.

***Guillaume Roels** is the Timken Chaired Professor of Global Technology and Innovation and Professor of Technology and Operations Management at*

INSEAD.

Hummy Song *is an Assistant Professor of Operations, Information and Decisions at the University of Pennsylvania's Wharton School.*

Mor Armony *is the Harvey Golub Professor of Business Leadership, a Professor of Technology, Operations, and Statistics, and the Vice Dean of Faculty at New York University's Stern School of Business.*

Don't miss our latest content. Download the free ***INSEAD Knowledge app*** *today.*

Follow INSEAD Knowledge on ***Twitter*** *and* ***Facebook****.*

Find article at

https://knowledge.insead.edu/operations/when-several-queues-are-better-one

About the author(s)

Guillaume Roels is the Timken Chaired Professor of Global Technology and Innovation at INSEAD and the Research Director of the **INSEAD-Wharton Alliance**.