# How Queueing Styles Affect Server Behavior

This study focuses on how servers behave in queueing systems. In basic queueing theory it is assumed that server speed is stochastic and independent of other factors in the system, but experimental evidence has shown that this is not necessarily true. Batt and Terwiesch (Working Paper) have shown that service time increases in a heavily loaded system, and Schultz (1999) demonstrated that low inventory production systems do not decrease worker speed. However, there are other factors which could change server speed and accuracy. This study looks to identify the presence of behavioral effects on the speed and accuracy of servers by simulating a supermarket environment using a behavioral experiment. There are three factors which this study will investigate for their effects on speed/accuracy. 1) *Queue type* describes the system in which the server is operating. This is either a single queue system with parallel servers or a series of parallel servers each with their own queue. 2) *Payment method* is the way servers are paid. They are either paid a flat rate (e.g. hourly wage) or for each item they process correctly (incentive pay). 3) *Group size* is the number of servers in system which will be marked as either a large group or a small group.

## Hypotheses and Theory

Managers have used single queue systems rather than parallel queue systems in numerous contexts such as delis and the DMV by taking a number (Martin, LA Times, 2002) or using a physical line like bookstores, the post office and grocery stores (Fantasia, Times Union, 2009). According to Fantasia, some managers believe using a single queue instead of multiple parallel queues leads to a shorter waiting time for customers. The theory behind this argument can be seen by comparing a join the shortest queue (JSQ) model to a single queue model. Consider the expected waiting time in a JSQ model with threshold jockeying being a lower bound to a general JSQ model. A jockeying threshold of one behaves the same as single queue (Adan 1994). This result holds only if servers work at the same speed regardless of queue type.

There are several factors that can influence server behavior depending on their queue type. Servers receive feedback about their working speed from the number of people in their queue. If servers work at different speeds, a shorter queue on average indicates a faster speed (Adan 1991). By having their own queue, each server in a group has a feedback loop where they can tell how fast they are working based on the number of people in their queue. However in a single, shared queue, that feedback no longer exists. A server cannot compare themself to other servers based on the size of their queue. Conversely, a server cannot be judged by others based on queue size. It is difficult to attribute the total speed of the line to any one server. In *Control Theory*, workers self-adjust their work according to a feedback loop (Klein 1989). Servers use a referent standard to compare their work with a goal. In the parallel queues setting, that standard is the length of other servers' queues, but it is difficult to set a similar visual goal in a single queue setting.

Another difference between single and parallel queues is the interdependence of the servers. In a parallel queue setting, each server has a workload they can reduce only by doing work. The interdependence of the servers is not entirely apparent. In a single queue setting, the workload is visibly interdependent (Bendoly 2009). In a group of servers, visually comparing work between servers is difficult, and it becomes easier to free-ride on the work of others. This

can cause *Social Loafing*. An experiment using rope pulling (Steiner 1972) found that in an effortful task with a group of people, there is an inverse relationship between the size of the group and individual effort. This result was verified with a clapping experiment and named Social Loafing (Latané 1979). Subsequent research was compiled and generalized Social Loafing across many different group settings (Karau 1993). The combination of Control Theory and Social Loafing leads to

*Hypothesis 1*: With payment type and group size fixed, individual speed in single queue systems is slower than individual speed in parallel queue systems.

There are some factors that can limit *Hypothesis 1*. The first is the *Köhler Effect* (Köhler 1926). According to the Köhler Effect, individuals perform *better* working in a group than they do independently because they want to impress other group members. This has been shown in athletic performance (Osborn 2012), but competitive forces are much stronger in college athletics than in typical server/queue settings. Another weakening factor of Social Loafing is providing feedback to individuals. When people receive feedback on their performance, they are less likely to loaf (Karau 1993). In a supermarket setting, servers can be given items-per-minute (IPM) scores based on their average speed of scanning items and voids per shift as a measure of accuracy (Sackett Zedeck and Fogli 1988). The experiment in this study . Another potential factor that could improve server speed/accuracy is their attitude about the single queue system. Workers comfortable with their environment have a psychological motive for improved work performance (Bitner 1992). Servers who are more comfortable in a single queue system because there is less pressure to process their own line will work better. In the experiment, the decision of the line style is made exogenously to help mitigate this work environment effect.

Managers must decide a payment method for servers. Paying servers for their performance can have a motivating influence on their performance. According to *Equity Theory*, servers are more motivated to work in a way that their effort is correlated with fair compensation (Donovan 2001). Payment for each correct job finished provides a motivation to work quickly and accurately. In contrast, paying a flat rate has no monetary benefit for working with a higher speed or accuracy. This leads to

*Hypothesis 2A*: With queue type and group size fixed, individual speed with flat payment is slower than individual speed with incentive pay.

*Hypothesis 2B*: With queue type and group size fixed, individual accuracy with flat payment is higher than individual accuracy with incentive pay.

Incentive pay should have a moderating impact on Social Loafing and Control Theory. According to *Expectancy Theory*, the motivational force to act is the product of the valence, instrumentality, and expectancy to the worker (Vroom 1964). A worker will move in the direction of the strongest positive force. Incentive pay increases the instrumentality of working hard under all situations. If servers have a large extrinsic motivation to perform well, that force should mitigate differences in the queue style treatment.

For a given customer arrival rate, managers must decide to use some number of servers. A common measurement used to determine the number of servers needed is *utilization*,  $\rho = \frac{\lambda}{K\mu}$ , where  $\lambda$  is the customer arrival rate,  $\mu$  is the average processing rate of one server, and *K* is the number of servers (Nelson 1989). By rearranging this equation based on a maximum desired utilization, a manager needs  $K = \left[\frac{\lambda}{\rho\mu}\right]$  servers where [x] is the ceiling function. Since  $\mu$  and  $\rho$  are not easily changed by management, the decision of how many servers to employ is based on  $\lambda$ . Therefore a manager cannot readily change the number of servers needed in order to reduce Social Loafing. Social Loafing effects are stronger with a larger group, therefore

*Hypothesis 3*: With queue type and payment type fixed, individual speed in a large group is slower than individual speed in a small group

If Social Loafing is mitigated through peer pressure or feedback, *Hypothesis 3* will be limited like *Hypothesis 1*. Also since Social Loafing occurs in single queue systems and large groups, combining these two factors should cause an interaction effect.

*Hypothesis 4*: There is an interaction effect between single-queue and large group that further decreases the speed of servers.

## Model and System Parameters

This paper will consider two models for queueing systems: a single-queue model with parallel servers and a JSQ model. The main result needed for comparison from each model is expected wait time, E[W], for a customer. Preliminary testing for the experiment found an average processing rate of  $\mu = 4$  customers / minute for with identical task descriptions.

#### **Single-Queue Model**

Consider a single queue with *K* parallel servers. Each server is identical and has an exponential service time of mean  $\mu_s$ . Customers arrive to the infinite capacity queue according to a Poisson process with mean  $\lambda$ . Let server utilization be  $\rho = \lambda/K\mu_s$ . Restrict  $0 < \rho < 1$  so that the system is ergodic.

The expected wait time for this model,  $E[W]_s$  can be found in Larson (1981). Since we are interested in comparing worker speed, we want the worker lines to be non-empty most of the time. For the experiment, the single-queue model is used as the baseline such that  $E[W]_s = 65$  seconds. Using the equations from Larson, a small group of K = 3 needs an arrival rate  $\lambda = 10.99$  customers / minute. This yields utilization  $\rho = 0.9155$ . For a large group of K = 8, we need an arrival rate of  $\lambda = 30.93$  customers / minute for utilization  $\rho = 0.9665$ .

#### JSQ Model

Consider a queue model with *K* systems. Each system has an identical server with exponential service time of mean  $\mu_m$  and a queue with infinite capacity. Customers arrive to the system according to a Poisson process with mean  $\lambda$  and join the shortest queue among *K*. Ties are broken uniformly. Let process utilization be  $\rho = \frac{\lambda}{\kappa\mu}$ . Restrict  $0 < \rho < 1$  so that the system is ergodic.

Due to the interdependent nature of the state space created by multiple queues, it is difficult to generate useful performance characteristics through mathematical analysis (Adan 1994). This means that numerical approximations are necessary for useful results. This paper will use the approximation method by Nelson and Philips (1989) for expected waiting time.

Nelson and Philips provide an error of less than one half of one percent if  $K \le 8$ . The maximum group size tested in our study is eight.

In order to compare the two models fairly, the same arrival rate is used as in the singlequeue model:  $\lambda = 10.99$  for K = 3 and  $\lambda = 30.93$  for K = 8. Applying Nelson and Philips' result to this data means the expected wait time  $E[W]_m = 75.02$  when K = 3 and  $E[W]_m = 74.25$  when K = 8.

#### Results

Since there are factors present which should slow down the servers, one must consider how much slower workers can work in the single-queue model before  $E[W]_m \le E[W]_s$ . For K =3, holding  $\mu_m = 4$  customers / minute, then any  $\mu_s < 3.974$  customers / minute would result in the JSQ model having a shorter expected waiting time. For K = 8 we need  $\mu_s < 3.981$  customers / minute.

### **Experiment Procedure**

The experiment takes place in an experimental lab at a medium-sized, private university in the Northeast United States with undergraduate management students. Participants are assigned a computer where they act as a cashier in a supermarket simulation. Each participant sees their queue of customers and the queue of computer-controlled servers. In the single-queue model, participants see the single queue. The task is to move a slider for each item in the customer's cart to the appropriate value. Sliders are a real effort task (Gill 2009). At the end of the experiment, each participant is given feedback on their performance. They are told how many items they processed in total and how many they processed correctly.

The factorial design of the experiment is 2x2x2. Group size (three or eight), queue type (parallel or single), and payment method (flat or incentive) make up the treatments.

Customers are automated so that they behave the same for each subject. Customer behavior is programmed with the following assumptions. Customers do not jockey if they see a shorter queue. When customers arrive to parallel queues, they will observe a random number of queues with each queue having an equal probability of being in the set of queues chosen. The customer joins the shortest queue in this set with ties settled randomly (see Graham 2000). A perfectly rational customer would join the line with the shortest expected wait time, but this would require the customer to know the speed of the servers. The customers do not adjust their lane choice by learning about the speed of servers, nor do they jockey when a shorter line is available. Customers have a Poisson arrival rate. They will have exactly five items in their cart. This cart size is used to make processing time approximately 15 seconds. A large cart would reduce the number of customers that each participant sees, but a smaller cart size would have the participant pushing the "Next" button too frequently. The computer-controlled servers process customers following an exponential distribution with a mean rate  $\mu = 4$  customers / minute.

Some model assumptions are made about the participants. Since group dynamics are an important part of this study, people who know each other in the lab could be less likely to engage in Social Loafing. To help create a group environment, a short team building exercise will begin the experiment in order to increase group cohesion for all groups. Group cohesion metrics will be taken in the exit survey. Social Loafing is present in undergraduate students as well as in the labor force, hence we expect their behavior to be representative (Karau 1993).

# **Preliminary Results**

At the time of this writing, results are still being gathered and analyzed with an expected completion date of April 30, 2013. Preliminary results are consistent with our hypotheses.

## References

- Adan, I. J. B. F., J. Wessels, W. H. M. Zijm. 1991. Analysis of the asymmetric shortest queue problem with threshold jockeying. *Stochastic Models* **7**(4) 615-627.
- Adan, I., van Houtum, G.-J., van der Wal, J. 1994. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research* **48** 197-217.
- Batt, R. J., C. Terwiesch. 2012. Doctors Under Load: An Empirical Study of State-Dependent Service Times in Emergency Care. Working Paper, Wharton.
- Bendoly, E., R. Croson, P. Goncalves, K. Schultz. 2009. Bodies of Knowledge for Research in Behavioral Operations. *Production and Operations Management* **19**(4) 434-452.
- Bitner, M. J. Servicescapes: The Impact of Physical Surroundings on Customers and Employees. *Journal of Marketing* **56** 57-71.
- Donovan, J. J. 2001. Work Motivation. Anderson, N., D. S. Ones, H. K. Sinangil, C. Viswesvaran, eds. *Handbook of Industrial, Work & Organizational Psychology*, V2. Sage Publications, London, 53-76.
- Fantasia, R. 2009. Move up to the head of the line. Times Union August 1, 2009. Albany, NY.
- Gill, D., V. Prowse. 2011. A Novel Computerized Real Effort Task Based on Sliders. Working Paper, University of Southampton.
- Graham, C. 2000. Chaoticity on Path Space for a Queueing Network with Selection of the Shortest Queue among Several. *Journal of Applied Probability* **37**(1) 198-211.
- Karau, S. J., K. D. Williams. 1993. Social Loafing: A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology* **65**(4) 681-706.
- Klein, H. J. 1989. An Integrated Control Theory Model of Work Motivation. *The Academy of Management Review* **14**(2) 150-172.
- Köhler, O. 1926. Kraftleistungen bei Einzel- und Gruppenarbeit (Strength performance in individual and group work). *Industrielle Psychotechnik* **3** 274–282.
- Larson, Richard C., Amadeo R. Odoni. 1981. Urban Operations Research. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
- Latané, B., K. Williams, S. Harkins. 1979. Many Hands Make Light Work: The Causes and Consequences of Social Loafing. *Journal of Personality and Social Psychology* **37**(6) 822-832.
- Martin, H. 2002. *Queuing Up for Different Kind of Wait at the DMV*. LA Times February 5, 2002.
- Nelson, Randolph D., Thomas K. Philips. 1989. An Approximation to the Response Time for Shortest Queue Routing. *Performance Evaluation Review* **17**(1) 181-189.
- Osborn, K. A., B. C.Irwin, N. J. Skogsberg, D. L. Feltz. 2012. The Köhler Effect: Motivation Gains and Losses in Real Sports Groups. *Sports, Exercise, and Performance Psychology* **1**(4) 242-253.
- Sackett, P. R. S. Zedeck, L. Fogli. 1988. Relations Between Measures of Typical and Maximum Job Performance. *Journal of Applied Psychology* **73**(3) 482-486.
- Schultz, K. L., D. C. Juran, J. W. Boudreau. 1999. The Effects of Low Inventory on the Development of Productivity Norms. *Management Science* **45**(12) 1664-1678.
- Steiner, I. D. 1972. Group process and productivity. San Diego, CA: Academic Press
- Vroom, V. H. 1964. Work and motivation. New York: John Wiley & Sons, Inc. 8-28.