Replication Report for "Humans Are Not Machines: The Behavioral Impact of Queueing Design on Service Time"

By: Masha Shunko, Julie Niederhoff, and Yaroslav Rosokha

Primary Team: Xiaoyang Long and Jordan Tong
University of Wisconsin-Madison

Secondary Team: Andrew M. Davis and Blair Flicker
Cornell University and University of South Carolina

Shunko et al. (2018) investigate the impact of queue design on worker productivity in service systems that involve human servers. The authors vary the queue structure (multiple parallel queues versus single pooled queue) as well as the visibility of the queue to the servers (visible or blocked). They find that the single-queue structure slows down servers, and that poor visibility of queue length slows down servers in absence of performance incentives.

Hypothesis to replicate:

Corresponding to Hypothesis 1 (Impact of Queue Structure), service times are shorter when customers are aligned into multiple parallel queues instead of a single pooled queue (when queues are visible and pay is flat).

Power Analysis

In the statistical analysis of the original study, the unit of observation is the median service time for the carts completed in the second half of the experiment (defined on p. 460) for each subject. By reconstruction using the original data posted by the authors, we determined that the definition of this variable is as follows: If a participant completed n carts, the authors used the median completion time from cart number $\lfloor n/2 \rfloor$ to number n. The authors then compare the 50th percentiles of median service time under different conditions. Specifically, for each comparison, they conduct a nonparametric permutation test with $10,000$ random permutations with replacement (see details of their approach on page 463 of the paper, and R code in the supplementary materials).

The original paper repeats the same experiment two times with different subject pools: once "in a behavioral lab at a large private university in the northeastern United States," and once on M-Turk. The authors preferred that the replication team replicate the original behavioral-lab experiment using a behavioral lab subject pool, but due to in-person restrictions from Covid-19 the replication team decided to replicate the original M-Turk experiment with an M-Turk subject pool.

In the original M-Turk experiment, the sample size is 113 subjects (53 for *parallel* and 60 for *single*). For the parallel vs. single queue conditions with M-Turk subjects under visible queue and flat pay (the two conditions we replicate), the paper reports that the difference in median service times is significant at the 5% level. The authors kindly provided the exact p-value as $p = 0.027$. Based on this reported p-value, we calculated that to achieve 90% power, we need at least 244 subjects, or 122 subjects per condition.

Sample

In the original M-Turk experiment, subjects were recruited from the pool of U.S.-based workers with at least 70% positive feedback and 50 successfully completed prior tasks. Subjects' ages ranged from 19 to 51, with a mean of 34.1 and a median of 33. We recruited M-Turk workers for both the primary and secondary replications using the same set of criteria as in the original study. The recruitment for the secondary replication occurred one week after the primary replication. The target sample size for each replication was 244 subjects.

Materials

Instructions were included in the appendices of the published paper. We used the same instructions. The authors kindly shared several JavaScript code files that contained the main logic of the queue simulator, as well as Qualtrics files containing the flow of the original experiment. We used these materials to recreate the experiment based on the descriptions in the published paper. In particular, we wrote new "outer shell" scripts (in HTML and CSS) to recreate the visuals of the experiment based on the screenshots in the published paper. Videos documenting the exact experiment process and stimuli we used are available online.[1]

Procedure

We follow the same protocols outlined in Section 3 of the paper.

Each subject works as a part of a group of four cashiers in which the three other cashiers are computer simulated. The cashiers are responsible for processing customers' carts; each cart contains five grocery items with different prices (ranging from $1 to $5). The subject's task is to move each of five sliders to a value corresponding to the price of each grocery item and then click the "Submit Cart" button. Computer-simulated customers arrive according to a Poisson process with a mean interarrival time of 5.5 seconds. The service time of each computer-simulated cashier is a random time set to 10 seconds plus an exponential random variable with a mean of 10 seconds.

Subjects are randomly assigned to one of two conditions: *single* and *parallel*. In the *single* condition, arriving customers join a single pooled queue. In the *parallel* condition, arriving customers join the shortest of multiple parallel queues, with ties broken randomly.

In the experiment, subjects are first shown a series of instruction screens that describe the experimental environment and task. They then complete a two-minute training session, after which they complete a 10-minute round of the experiment. Each subject is paid a flat fee of $3 for their participation. The pre-registration report for the experiment is available at https://aspredicted.org/yq3sk.pdf.

Analysis

We apply the same analysis technique as in the original paper by adapting the R code provided in the original article's supplementary materials to accommodate the fact that we only conducted two conditions. As discussed in the Power Analysis section,

[1] See https://osf.io/ua8kw/?view_only=09cca4f6f2d344d8a546f4e3f01757b2.

the primary analysis consists of comparing the medians of the dependent variable between the two conditions by conducting a nonparametric permutation test with 10,000 random permutations with replacement. (Note that estimates may vary in successive R code executions due to the bootstrapping procedure.)

As secondary analysis, we perform linear regression predicting the dependent variable with a dummy variable indicating whether the queue is single or parallel, and controlling for whether the subject has managerial experience, the device used, gender and age. We follow the "robust regression approach and find MM-estimators using the robustbase package in R" as described in subsection 5.3 in the paper.

Differences from Original Study

To adhere to institutional minimum payment standards at the time of replication, we pay subjects $3 for participating in the experiment, rather than the original amount of $1.25. There may also be some minor differences in the look of the experiment due to the new "outer shell" scripts written based only on the static screenshots available in the published paper, or in the adaptation of code files in order to make the experiment function as described in the original paper.[2]

Replication Results

329 subjects completed the study for each of the primary and secondary replications on M-Turk. After consulting with the authors, we decided to exclude subjects who completed 0 carts because there are no such observations in the original posted data. After these and other preregistered exclusions, 246 and 252 subjects were included in the analysis for primary and secondary replications, respectively. See Table 1 for exclusions by location, condition, and causes, as well as the resulting number of subjects included in the analysis.

Table 2 shows the results of the primary analysis that executes the nonparametric permutation test to compare conditions, along with the corresponding results reported in Shunko et al. (2018). For the primary replication (Wisconsin), the difference between conditions is directionally consistent with the hypothesis (21.72 in *parallel* vs. 21.94 in *single*). The difference is not statistically significant at the $p < 0.05$ level ($p = 0.444$). For the secondary replication (USC), the difference between conditions is not directionally consistent with the hypothesis (24.60 in *parallel* vs. 24.24 in *single*). The difference between conditions is not statistically significant at the $p < 0.05$ level ($p = 0.430$).

Table 3 shows the results of the secondary analysis that implements a robust regression approach, along with the corresponding coefficient estimates reported in Shunko et al. (2018). For both primary and secondary replications, the coefficient corresponding to queue configuration is not statistically significant at the $p < 0.05$ level.

Unplanned Protocol Deviations

There were no unplanned protocol deviations.

Discussion

In summary, across both replications on M-Turk, we did not find a significant difference at the $p < 0.05$ level between the second-half median service times under the *single* queue configuration relative to that under

[2] For instance, we wrote new code that checks for cart accuracy, which is described in the original paper but absent from the files shared with us.

the *parallel* queue configuration. The difference between conditions was directionally consistent with the hypothesis for the primary replication, but not for the secondary replication. Robust regression analysis also did not result in a statistically significant difference between *single* and *parallel* conditions at the $p < 0.05$ level.

We note that although we have adhered to the same selection criteria in the replication experiments as those in the original experiments, there is evidence of differences between the two subject pools. First, the second-half median service times were larger in the replication studies (ranging from 21.72 to 24.60) than those reported in Shunko et al. (2018) (ranging from 15.63 to 18.00). We note that these patterns suggest that queues were generally longer in the replications than in the original experiment. Second, for non-manipulated control variables (e.g., whether the subject is Male), we observe different coefficients (see Table 3).

Table 1 Number of subjects included in analysis, with exclusions by cause, condition, and replication

	Single			*Parallel*		
Experiment	Excluded 0 carts	Excluded other	Included in analysis	Excluded 0 carts	Excluded other	Included in Analysis
SNR (2018)	NR	NR	53	NR	NR	60
Primary M-Turk Rep. (Wisc.)	36	7	124	27	13	122
Secondary M-Turk Rep. (USC)	41	1	122	34	1	130

Notes. "Other" exclusions are all technical such as not being able to locate payment record or duplicate IP address. "NR" indicates that these numbers were not reported in the original manuscript.

Table 2 50th percentiles of second-half median service times, with condition comparisons

Experiment	*Parallel*	*Single*	p-value
SNR (2018)	15.63 (0.75)	18.00 (0.97)	0.027
Primary M-Turk Rep. (Wisc.)	21.72 (1.10)	21.94 (1.34)	0.444
Secondary M-Turk Rep. (USC)	24.60 (1.67)	24.24 (0.98)	0.430

Note. Bootstrapped standard errors reported in parentheses.

References

Shunko, Masha, Julie Niederhoff, Yaroslav Rosokha. 2018. Humans are not machines: The behavioral impact of queueing design on service time. *Management Science* **64**(1) 453–473.

Table 3 **Regression results for replications and original experiment**

	SNR (2018)		Primary		Secondary	
Variable	Estimate	p-value	Estimate	p-value	Estimate	p-value
Constant	17.883		17.401		25.552	
Parallel	-1.530	0.000	0.026	0.982	0.225	0.877
	(0.371)		(1.171)		(1.457)	
Born≥1990	-2.156	0.000	-0.406	0.749	0.312	0.844
	(0.437)		(1.267)		(1.585)	
Male	-1.707	0.000	2.564	0.035	-2.347	0.134
	(0.374)		(1.209)		(1.56)	
Managerial	0.512	0.208	3.687	0.001	0.522	0.735
	(0.406)		(1.125)		(1.54)	
TouchPad	1.746	0.000	4.909	0.001	0.747	0.636
	(0.388)		(1.454)		(1.575)	
TouchScreen	2.121	0.049	-1.764	0.220	-2.827	0.447
	(1.073)		(1.435)		(3.715)	

Note: Standard errors reported in parentheses. Note that the estimates reported from the regression table in Shunko et al. (2018) included other variables for conditions not run in the replications.