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

by Masha Shunko, Julie Niederhoff, and Yaroslav Rosokha
June 11, 2022

We applaud Davis et al.'s (2022) efforts to replicate the results of one of our hypotheses presented in Shunko et al. (2018). We strongly believe in the result that "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)," which, in addition to being documented in our study, has been observed independently in experimental (Song et al. 2015 and, with some caveats, Song et al. 2021), and empirical (Wang and Zhou 2018) studies. It has also been well accepted in operations management literature that accounts for behavioral effects (Rothkopf and Rech 1987; Armony et al. 2021; Ping et al. 2021; Sunar et al. 2021). However, despite their efforts, Davis et al. (2022) were unable to replicate our hypothesis. Given independent evidence from both laboratory and field experiments, we have no doubt that our key result holds, but perhaps Davis et al. (2022) identified an important boundary condition, which future literature can scrutinize.

Before we address the difference in the results, we want to point out that there were some hidden differences in the experimental conditions, which led to, as Davis et al. (2022) noted, significant differences in the coefficients of the control variables, as shown in Table 3. For instance, the coefficient for *Born > 1990* was negative and significant in our study (-2.156, p-value < 0.001), but it was insignificant in Davis et al.'s (2022) study. The *Male* coefficient was negative and significant in our study (-1.707, p-value <0.001), but it was positive and significant in the primary replication in Davis et al. (2022) (2.564, p-value = .035) and insignificant in the secondary replication. *TouchPad* and *TouchScreen* were positive and significant in our study (1.746, p-value <0.001, and 2.121, p-value = 0.049 respectively), but they were mostly insignificant in Davis et al. (2022). These differences in the controls seem to point to something fundamentally different in the replication study, and the coefficients in Davis et al. (2022) do not make intuitive sense to us. Relatedly, we are puzzled by why the control variables across two replications in Davis et al. (2022) are very different in terms of signs and significance. For example, the *Male* coefficient in the primary replication is positive and significant (2.564, p-value = .035), while in the secondary replication it is negative and insignificant (-2.347, p-value = .134); the *Managerial* coefficient in the primary replication is positive and

significant (3.687, p-value = .001), while it is insignificant (p-value = .735) in the secondary replication; and the *TouchPad* coefficient in the primary replication is positive and significant (4.909, p-value = .001), while in the secondary replication it is insignificant (p-value = 0.636). These differences seem to indicate that, across even two replications, the experimental results lack consistency, unlike our results, which were consistent across the laboratory and the M-Turk. We propose that these differences arose because the experimental conditions had changed since 2013–2014, when we conducted our experiments: the pool of M-Turk participants may have become more diverse, the devices the M-Turk workers used to complete the experiment may have changed (e.g., smartphones are more popular and widespread now, and the experiment is much harder to complete on a phone compared to a computer), and the M-Turk interface was redesigned in 2017 (Amazon Mechanical Turk Blog 2018). All of these issues could have led to different processing times and increased worker performance variability.

Are these differences consequential? We believe so. Namely, as Davis et al. (2022) noted, the median cart submission time in our paper was 15–18 seconds across all treatments, both on M-Turk and in the lab. However, in Davis et al. (2022), the median cart submission time was closer to 22–25 seconds. This is a significant increase (almost 50%), which is not easy to explain. Moreover, one of the SNR coauthors of this note used the same real-effort cart submission task created with a different code and without queue visualization. They observed cart submission times of around 17–18 seconds (see Table 8 in Choudhary et al. Forthcoming), which were similar to the times obtained in the original Shunko et al. (2018) paper. This independent evidence gives us further confidence in our results.

Does cart submission time matter? We believe so because an important consequence of the significantly slower submission times is that the queue would become more congested, leading to *longer queues* than the queues in our experiment in both parallel and single queue settings. This is consequential because behavioral literature suggests that performance is highly impacted by the system load (see, for example, Schultz et al. 1998; Tan and Netessine 2014); hence, we may expect different behaviors and results in replication experiments with higher waiting times. In the original paper, we ran robustness tests with low loads, but Davis et al.'s (2022) setting was the opposite condition, which we did not study.

While we can only speculate why there are important differences in the results of our study and Davis et al.'s (2022) study, it seems that future work should attempt to replicate and compare results under high and low workloads. It appears that our key result should continue to hold under low workloads, but it may disappear under higher workloads. However, given that the relationship between workload and productivity is nonlinear (Tan and Netessine 2014), the relationship might be more nuanced. We hope that future research can address this question, and we thank Davis et al. (2022) for potentially identifying this interesting boundary condition.

References

1. Davis A. M., Flicker B., Long X. and Tong J. 2022. Replication Report for Humans Are Not Machines: The Behavioral Impact of Queueing Design on Service Time"
2. Shunko M, Niederhoff J. and Rosokha, Y. 2018. Humans are not machines: The behavioral impact of queueing design on service time. *Management Science* 64(1) 453–473.
3. Wang J and Zhou YP. 2018. Impact of queue configuration on service time: evidence from a supermarket. *Management Science* 64(7), 3055–3075.
4. Armony M, Roles G, and Song H. 2021. Pooling queues with strategic servers: The effects of customer ownership. *Operations Research* 69(1), 13–29.
5. Rothkopf MH and Rech P. 1987. Perspectives on queues: combining queues is not always beneficial. *Operations Research* 35, 906–909.
6. Sunar N, Tu Y, and Ziya S. 2021. Pooled vs. dedicated queues when customers are delay-sensitive. *Management Science* 67(6), 3785–3802.
7. Song H, Tucker AL, and Murrell KL. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12), 3032–3053.
8. Schultz KL, Juran DC, Boudreau JW, McClain JO, and Thomas LJ. 1998. Modeling and worker motivation in JIT production systems. *Management Science* 45(12), 1595–1607.
9. Song H, Armony M, and Roels G. 2021. *Queue configuration and server's customer orientation: An experimental investigation*. INSEAD Working Paper No. 2021/71/TOM/ACGRE, Available at SSRN: https://ssrn.com/abstract=3980495 or http://dx.doi.org/10.2139/ssrn.3980495 SSRN 3980495
10. Choudhary V, Shunko M, Netessine S, and Koo S. (Forthcoming). Nudging drivers to safety: Evidence from a field experiment. *Management Science*.
11. Tan T. and Netessine S. 2014. When does the devil make work? An empirical study of the impact of workload on worker performance. *Management Science* 60(6), 1574–1593.
12. A Worker site improvement to help increase Worker productivity. 2018. *Amazon Mechanical Turk*, June 10, 2018 [Blog]. Available at https://blog.mturk.com/a-worker-site-improvement-to-help-increase-worker-productivity-3b882f2bc78f (Accessed Month June, 2022).