Online Appendix for "A Replication Study of Operations Management Experiments in *Management Science*"

A. Paper-Selection Survey Details

As described in Section 2, the paper-selection survey was announced via post to the MSOM Society and Behavioral Operations Management sections of INFORMS. Elena Katok posted an initial announcement on September 4, 2020, entitled "Management Science Replication Project," and Andrew Davis posted a reminder on September 29 entitled "Last Call – Management Science Replication Project."

Survey respondents were given the following instructions: "For each sub-area, we ask you to identify up to two papers that you are most interested in seeing replicated. Once the survey is completed, for each sub-area, we will select the paper with the highest vote total. Additionally, we will also select between three to five other papers with the highest vote totals across all sub-areas. This will ensure coverage across all key areas of operations management, while also including additional papers that are deemed especially important to replicate."

Respondents could vote for zero, one, or two papers in each category. Responses that did not register a vote for any paper in any category was deemed invalid. Between September 4 and October 7, we received 97 valid responses. The categories, papers, and descriptions provided to survey respondents appear below. Additionally, the number of votes each received is noted. We included the two papers from each category that received the highest number of votes (marked with an arrow).

The survey ended with two contextual questions. The first asked respondents to rate the importance of various factors in selecting their responses on a 5-point Likert scale from "Not important" (score of 1) to "Very important" (score of 5). The items (with mean scores in parentheses) were: familiarity with the papers (2.9); interest in the application area (3.7); importance to the field of OM (4.1); and likelihood of successful replication (2.3). The second provided a list of eleven application areas (behavioral operations, contracting, forecasting, healthcare, inventory management, new product development, procurement, retailing, revenue management and pricing, service operations, and sustainable operations) and asked respondents to check all in which they conduct research. 47% of the responses included "behavioral operations" in their choice set.

Inventory Management

↦ **Chen et al. (2013), 38 votes**: In the newsvendor setting, order-time payments receive less weight than demand-time payments.

Ho et al. (2010), 29 votes: The "pull-to-center" bias exists in a multilocation newsvendor setting and the effect is stronger in a high-profit environment.

Lee et al. (2018), 25 votes: Task decomposition improves newsvendor order quantities in a high-profit environment.

Ren and Croson (2013), 30 votes: Overprecision correlates with the newsvendor "pull-to-center" bias.

↦ **Schweitzer and Cachon (2000), 49 votes**: Newsvendor order quantities exhibit a "pull-to-center" bias (i.e., quantities are set between normative predictions and mean demand).

Supply Chain and Contracting

Becker-Peth et al. (2013), 20 votes: In buyback contracts, specific parameters influence order quantities, not only the critical ratio. A behaviorally optimized contract achieves better efficiency.

Beer et al. (2018), 33 votes: Reciprocal intentions can be signaled by means of a relationship-specific investment, which leads to more collaborative supply chain relationships.

↦ **Croson and Donohue (2006), 37 votes**: A bullwhip effect in supply chain order quantities occurs even when the demand distribution is known and stationary, but sharing inventory information decreases the effect.

↦ **Ho and Zhang (2008), 36 votes**: In pricing contracts, it is more efficient to frame a fixed-fee as a quantity discount, than as a two-part tariff.

Katok and Wu (2009), 35 votes: Revenue sharing and buyback contracts are mathematically equivalent, but experimental results show their different framings generate different supply chain performance.

Queuing and Production

Buell and Norton (2011), 40 votes: "Operational transparency" (i.e. seeing the work being done) affects customer value in a simulated airline price search website. Perceived customer value is higher in the transparency condition (holding customer wait fixed).

↦ **Kremer and Debo (2016), 50 votes**: Uninformed consumers can infer quality from observed wait times in a simulated queuing experiment if there are enough informed consumers—purchase probability increases in the share of informed consumers.

Schultz et al. (1998), 23 votes: The"Independence Assumption" (i.e. that worker productivity does not depend on others' speed or the WIP state) fails in a data entry task. In the Low Inventory setting worker productivity increases when they are slower than and/or delaying other workers.

↦ **Shunko et al. (2018), 55 votes**: When queues are visible, service times are shorter in parallel (i.e. unpooled) queues compared to single (i.e. pooled) queues.

Forecasting

Feiler et al. (2013), 28 votes: Beliefs about the true demand mean are biased towards the observed censored sales data.

↦ **Kremer et al. (2011), 37 votes**: Forecasters overreact to forecast errors in relatively stable environments, but underreact to forecast errors in relative unstable environments.

Kremer et al. (2016), 19 votes: Direct-top forecasts are better (worse) than bottom-up forecasts when item-level demand noise terms are negatively (positively) correlated.

↦ **Özer et al. (2011), 45 votes**: The manufacturer's report is positively correlated with their private information; the supplier's decision is positively correlated with the manufacturer's forecast.

Scheele et al. (2018), 32 votes: Forecasters are averse to lying and overreact to forecast error penalties.

Sourcing & Procurement

Chen-Ritzo et al. (2005), 24 votes: An ascending multiattribute (price, quality and lead time) auction mechanism increases both buyer utility and supplier profits compared to a price-only auction.

↦ **Davis et al. (2011), 37 votes**: Sellers deviate from theoretical predictions when setting reserve prices in a matter consistent with risk aversion, anticipated regret, and probability weighting.

↦ **Engelbrecht-Wiggans and Katok (2008), 37 votes**: Winner's regret (winning an auction and paying too much) results in lower average bids; loser's regret (missing opportunities to win at a favorable price) results in higher bids.

Fugger et al. (2016), 20 votes: Dynamic nonbinding reverse auctions enable suppliers to collude, leading to noncompetitive prices.

Kwasnica et al. (2005), 21 votes: For auctions over many objects, the Resource Allocation Design (RAD) auction, an iterative auction with package bidding, outperforms the Simultaneous Multiple Round auction (SMR) used by the FCC: RAD achieves higher efficiency, higher net revenue, lower bidder losses, and fewer iterations.

B. Prediction Survey Details

The prediction survey was developed to elicit respondents' predictions for the probability of each paper replicating. We modeled our survey on those used by Dreber et al. (2015), Camerer et al. (2016), and DellaVigna et al. (2019). Because we were replicating each paper at two separate sites, we asked respondents to predict the likelihood of a "full replication", defined as: *the replication data finds statistically significant results ($p < 0.05$) in the same direction as the original study at both sites (evaluated separately). Studies that have an in-person follow-up replication to an initial online non-replication will be deemed to have replicated if the in-person follow-up finds statistically significant results. If the finding to replicate involves more than one statistical test, all of the tests must be statistically significant to qualify as a full replication.*

An initial page described the context and replication procedures for the project. Each paper was then presented with a summary of the "finding to replicate" and additional information such as the original sample size and p-value, the replication sample size and statistical power, and the replication modality. Links were also provided to the original paper and the preliminary replication report. Respondents were asked to predict the likelihood of replication (ranging from 0% to 100%), and their confidence in their prediction (on a 7 point scale ranging from "No Confidence" to "Very High Confidence"). Figure B.1 shows an example elicitation for one paper.

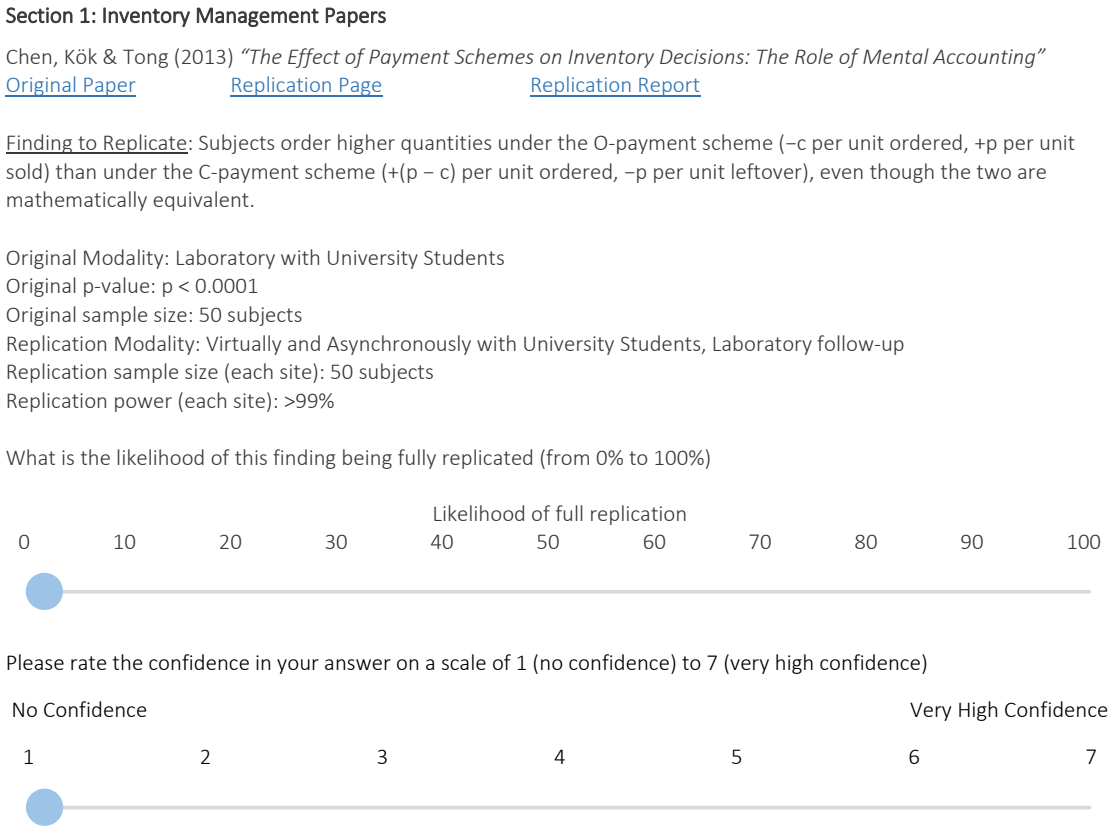

Figure B.1 Example Prediction Elicitation

The survey finished with several questions about the respondent. Respondents were asked to identify in which application areas they conducted research, which included "Behavioral operations" (alongside topics such as "Healthcare", "Inventory management", "Service operations", etc.). They were also asked whether they had ever published a paper with a laboratory experiment. These questions were designed to allow us to examine whether predictions differed systematically between researchers focused on behavioral and experimental research versus the broader OM community. Finally, respondents were asked how familiar they were with the papers (on a 7 point scale ranging from "Not at all familiar" to "Very familiar") and whether they were an author of one of the papers.

Announcements were then sent to the MSOM Society and the Behavioral Operations Management Section of INFORMS on August 11, 2021, with follow-up reminders on September 1 and September 5. We closed the survey on September 17, 2021, after receiving 43 unique completed responses. 21 respondents identified "Behavioral operations" as an area of focus. 15 repondendents reported having published a paper with a laboratory experiment. Three respondents identified as authors of a replication study.

C. Example of Initial Correspondence with Original Authors

Dear [**Authors**],

We are writing on behalf of the Management Science Replication Project team (see announcement here). In a recent survey that we sent out to the OM community (link to survey: `msreplication.com`), we asked people to identify the papers that they were most interested in seeing replicated. Your paper, "[**Paper Title**]," was voted as one of the top ten papers. This shows that the community believes that your paper is one of the most important/interesting experimental papers in behavioral OM.

As a next step, our team plans to replicate each selected study with relatively large samples, at two different universities, using these universities' participant pools. We will first conduct these replications online (**if an individual decision**). We will then attempt to replicate studies that do not replicate online in physical labs. [**Primary Team Member**] (cced) and I have been assigned to lead the replication effort for your paper. The two locations that are tentatively scheduled to run the experiment are [**Site 1**] and [**Site 2**].

To help us better replicate your experiment, we were wondering if you could kindly assist us with the following questions:

1. The key result we plan to test is: "[**Hypothesis**]" Do you think this accurately reflects the key result of your paper?
2. Do you have any details/materials that you used to run your experiments, and if so, would you be willing to share them with us? Doing so would help us replicate the experiment as closely as possible.
3. Do you have any other comments or suggestions that you think would be helpful in running the experiments?

Thank you for any help that you are willing to provide! If possible, it would be great if you could get back to us before [**Date**]. We will also keep you informed of the detailed replication plan once it is ready and will seek your feedback along the way.

Please let us know if you have any questions.

Sincerely,

[**Primary Team Members**]

D. Example of Correspondence with Original Authors Regarding the Initial Draft of the Replication Report

Dear [**Authors**],

Thank you so much for your help so far with the MS Replication Project for your paper: "[**Paper Title**]" Attached, you'll see that we have drafted a replication procedure plan and pre-registration document for the protocol. Could you please let us know if you have recommended edits/questions/concerns regarding these two documents by [**Date**]? Feel free to make small suggestions as pdf comments or in an email. If you have larger concerns that require a call, we are also happy to do so.

We are still finalizing the actual experimental interface and will seek your feedback before we run the experiment as well.

Thanks again!

Best,

[**Primary Team Members**]

E. Example of Correspondence with Original Authors Regarding the Final Replication Report, Data, and Analysis

Dear [**Authors**],

We hope this email finds you doing well! The Management Science Replication Project team has completed a draft of the replication report for your paper entitled "[**Paper Title**]." Specifically, attached to this email you will find a PDF of the replication report for your project as well as the data and code for the replications.

We are emailing because we want to provide you the opportunity to:

1. Review the report and provide the team with any feedback/suggestions,
2. Prepare a response document that we will post alongside the replication report, and
3. Ask any questions of us.

All the above opportunities are optional. If you do wish to take us up on them, though, there are some deadlines that we need to follow that we want to communicate with you:

- [**Date**] (~ 4 weeks): Deadline for suggestions/feedback on the replication report itself.
- [**Date**] (~ 8 weeks): Deadline for (optional) submission of a response document for us to post alongside the replication report.

Thank you for your attention, and please reach out if you have any questions or concerns.

Sincerely,

[**Primary Team Members**]

F. Responses from Original Authors

All original author teams were given the opportunity to draft a response document for us to post alongside the replication report online (see Appendix E). At the time of submission, we had only received responses from three teams of original authors: Croson and Donohue (2006, partial replication, see page 10), Ho and Zhang (2008, non-replication, see page 13), and Shunko et al. (2018, non-replication, see page 22). With their permission, we have also included the authors' responses on the following pages.

F.1. Croson and Donohue (2006)

Response to Replication Report for "Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information"

Rachel Croson

McKnight Endowed Professor of Economics,
Executive Vice President and Provost, University of Minnesota
rcroson@umn.edu

Karen Donohue

Professor and Curtis L. Carlson Chair in Supply Chain,
Carlson School of Management, University of Minnesota
donoh008@umn.edu

We would like to thank the Management Science Replication Project (MSRP) team for conducting this important study (Davis et al. 2022). We are strong advocates of replication-based research, both in principle, and as a way to deepen our knowledge of why, when, and how a behavioral effect of interest emerges. Identifying that an effect exists within a single setting should propel further inquiry rather than being the last word. The MSRP team provides a great service to the behavioral operations community in delivering this impressive set of replication studies, and we encourage everyone in our profession to support this important work, especially encouraging more editors to support and publish replication research. This direction will help ignite further inquiry into behavioral phenomena that is critical to improving how operations and supply chains function. The methodological rigor shown in this particular effort also serves as an aspirational benchmark for how experimental work should be conducted and is sure to elevate the research of future behavioral operations scholars.

When we began studying the bullwhip effect within the context of the classic Beer Distribution Game (Sterman 1992) over 20 years ago, one of the most surprising results was that the effect remains even when the distribution of demand is commonly known to all supply chain members. The base-case setting of the original paper (Croson and Donohue 2006) controlled for other well-known operational causes of the bullwhip effect (Lee et al. 1997), isolating the role of behavioral factors as important contributors to the bullwhip effect and confirming supply line underweighting (Sterman 1989) as one significant factor. These results, which are featured in Study 1 of our paper, set the stage for Study 2 which looked further at a possible intervention for reducing this form of bullwhip behavior; specifically sharing inventory information across the supply chain.

The MSRP team's work focuses on replicating the finding that "sharing inventory information with supply chain members helps alleviate, but not eliminate, the bullwhip effect" (i.e., Hypothesis 3 within Study 2). To do this, they conduct experiments at two sites: UTD and UW-Madison. The results are mixed, with the UW-Madison sample providing stronger support of Hypothesis 3 than the UTD sample. Both samples also show a wider range of order variation, with a "non-trivial number of apparent outliers" (Katok et al. 2022, page 3) compared with the sample generated through our original experiment.

These results are interesting in themselves, especially in comparing the UTD and UW-Madison samples. The results also raise other interesting questions in comparison to the original study. We offer three questions that we think are worthy of further exploration.

The first question is *what is the role of physical presence in creating common knowledge?* One difference in protocol between the UTD and UW-Madison samples was the physical location of the participants. In the UTD sample, students took part in the experiment in a hybrid mode with some in class and others connected from home. In the UW-Madison sample, as in our original experiment, students took part in the experiment together in a classroom. When participants are together in a physical setting, common knowledge is straightforward to create and obtain. Could this difference in physical proximity be a contributing factor to the differences seen in the UTD sample (which did not replicate), versus the UW-Madison and original samples (which did)? Physical proximity may reassure participants about the knowledge, attention and perceived capabilities of other players, something that may be missing when participants engage virtually. This finding and observation raises interesting methodological issues about how we run experiments. Perhaps more importantly, it raises substantive issues about the influence of proximity on decision making; issues that will gain more salience given changing trends toward working from home within the supply chain profession.

The second question is *how should outliers in experimental data be interpreted?* Both the UTD and UW-Madison samples contained a significant number of extreme outliers, compared to none in our original sample. One reason for this difference could be the larger size of the replication samples (roughly 8 times as large as the original sample), which increases the likelihood of drawing in individuals or groups with unusual strategies. While the main conclusions of both replication samples appear to be robust to various approaches to removing outliers, the existence of outliers also raises interesting issues about the volatility of this decision setting. We know that the Beer Distribution Game has a fragile equilibrium which can be hard to converge upon or reestablish once bullwhip behavior is triggered (Croson et al. 2014). Most behavioral research has examined what causes the bullwhip effect or what conditions (such as sharing

inventory information, in the case of our original study) helps to alleviate its magnitude. Evidence of extreme outliers suggests it would be helpful to also explore what interventions are effective when the bullwhip has already taken hold and supply chains seek to return to more stable ordering behavior. Such "out of equilibrium" settings are becoming the new normal, as we deal with uncertainty caused by global supply shortages, shipping delays, and swings in economic conditions.

The final question is *whether individual knowledge and experience of supply chain dynamics has changed in a way that leads to different outcomes in the Beer Distribution Game setting?* Our original study was conducted over 20 years ago, at a time when the bullwhip effect was less well known to our student population or the broader public. Today, a vast majority of students have learned about supply chain dynamics in their curriculum, and have recently witnessed order volatility first-hand, especially with the rollercoaster of demand shifts during the pandemic. While this goes well beyond the implications of this replication study, revisiting this research in today's environment allows us to identify the impact of this additional information and experience on behavior.

These questions were, and hopefully the answers will be, directly inspired by this replication. We look forward to continuing this conversation, and want to reiterate our support for replication research in general, and this initiative (and this project) in particular. It is only through replication, information and data-sharing that our science progresses, and this initiative is pivotal to that progress.

References

Croson, R. and Donohue, K. 2016. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science,* 52, 323-336.

Croson, R., Donohue, K., Katok, E. and Sterman, J. 2014. Order stability in supply chains: Coordination risk and the role of coordination stock. *Production and Operations Management*, 23, 176-196.

Davis, A., Flicker, B., Hyndman, K., Katok E., Keppler, S., Leider, S., Long, X., and Tong, J., 2022. A replication study of operations management experiments in *Management Science*, to appear in *Management Science.*

Katok, E., Hyndman, K., Tong, T., and Long, X. 2022. Replication report for "Behavioral Causes of the Bullwhip Effect and the Observed Value of Inventory Information", online appendix: A replication study of operations management experiments in *Management Science*.

Lee, H., Padmanbhan, P. and Wang, S. 1997. Information distribution in a supply chain: The bullwhip effect, *Management Science*, 43, 546-558.

Sterman, J. 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science,* 35, 321-339.

Sterman, J. 1992. Teaching takes off: Flight simulators for management education. OR/MS Today, 19(5), 40-44.

F.2. Ho and Zhang (2008)

Author Response to Replication of Ho and Zhang (2008)

Teck-Hua Ho and Juanjuan Zhang

June 11, 2022

We are honored to have been selected as one of ten *Management Science* studies to be replicated. We are grateful to the replication team for being transparent in their efforts and for sharing their data, code, and report (received on April 6, 2022).

Main Results of the Original Paper Were Replicated

Our original study tested alternative pricing-contract solutions to the inefficiency problem in the channel literature. Linear pricing contracts are inefficient because they lead to higher wholesale and retail prices than those that maximize total channel profit.

In standard economic theory, nonlinear pricing contracts offer a solution (Tirole 1988). For example, a two-part tariff (TPT) contract can restore channel efficiency. With a TPT contract, the manufacturer charges a fixed fee in addition to a wholesale price. This allows the manufacturer to lower the wholesale price, which induces the retailer to lower the retail price.

The main results of the original paper are as follows:

1. TPT does not solve the channel inefficiency problem because of loss aversion. Since the retailer is averse to paying a high fixed fee (an upfront loss), the manufacturer is forced to raise the wholesale price above what would otherwise have been offered without loss aversion, in order to maximize profit. The retailer responds by charging a higher retail price, to the detriment of channel efficiency.

2. Framing the TPT as a quantity discount (QD) improves channel efficiency even though the two contracts are economically equivalent. QD improves channel efficiency because it mitigates loss aversion by making it less salient. This allows the manufacturer to charge a higher fixed fee and a lower wholesale price, which induces the retailer to charge a lower retail price.

Both main results from the original paper were replicated (see the table on next page):

1. TPT was inefficient; in fact, its efficiency was even lower than in the original study.

2. Compared with TPT, QD produced a lower wholesale price, a higher fixed fee, and a lower retail price. All these effects were highly significant at the 0.001 level.

	Original Study			Replication Study		
Variable	TPT $N = 264$	QD $N = 242$	p-value	TPT $N = 528$	QD $N = 517$	p-value
Wholesale price	3.96 (1.17)	3.41 (1.25)	0.000	4.54 (1.45)	4.23 (1.66)	0.001
Fixed fee	5.24 (2.32)	6.95 (4.17)	0.000	4.34 (2.49)	6.02 (7.65)	0.000
Retail price (if accept)	6.86 (0.54)	6.71 (0.80)	0.029	7.06 (1.09)	6.80 (1.06)	0.001
Acceptance (%)	74.24 (43.81)	82.23 (38.30)	0.029	76.14 (42.67)	70.41 (45.69)	0.036
Efficiency (%)	69.51 (41.27)	76.37 (36.18)	0.047	65.18 (21.82)	62.62 (17.56)	0.328

Notes. Values in parentheses are standard deviations.

The only channel decision not replicated was the retailer's acceptance of the manufacturer's offer. The acceptance rate was lower in the QD condition of the replication study. Since the efficiency of a rejected contract was 0%, the overall channel efficiency became statistically indistinguishable between the TPT and QD conditions, contrary to the original study.

Note the following about the retailer acceptance rate:

1. The retailer acceptance rate is secondary to the main results of the original paper. As explained above, what is central to channel efficiency is charging lower wholesale and retail prices. To understand the logic, consider a manufacturer who charges a wholesale price equal to the marginal cost of production. This efficient contract may end up being rejected if the manufacturer leaves the retailer with too little surplus by charging a high fixed fee. As modeled in our original paper, this could happen if the manufacturer cannot perfectly predict the retailer's acceptance decision.

2. In the replication study, the acceptance rate's p-value of 0.036 was much higher than those of the other three channel decisions, which were 0.001 or less. In addition, there were observations in the replication data where the manufacturer's offer was such that the maximum profit the retailer could have earned was negative. These manufacturer-subjects probably did not understand the experimental task. If we were to remove these observations from the replication data, the difference in acceptance rate between the TPT and QD conditions would have been statistically insignificant ($p = 0.174$), whereas the statistical significance of the other three channel decisions would have remained to be 0.001 or less.

3. The replication study reported participant frustration and implemented unplanned protocol deviations that might have affected the acceptance rate and lowered channel efficiency in the QD condition. We address this in the next section.

Participant Frustration and Unplanned Protocol Deviations in the Replication Study

In the replication study, participants had unanticipated difficulty in completing the experimental task, especially in the QD condition. As acknowledged in the "Unplanned Protocol Deviations" section of the replication report, participants were frustrated; those in the TPT condition took two hours to complete the task while those in the QD condition needed substantially longer time.

In an e-mail to us, the replication team shared the following observation:

> *The reason QD took longer is because of extra calculations required for Price A. This calculation involves X + Y/(10-P) which is often not an integer. To do this calculation correctly also requires understanding order of operations, which to my great surprise it turns out that many people don't understand. So some tried to calculate it as (X+Y)/(10-P). Even after calculating Price A correctly, they often ended up with a fraction and then had to multiply a fraction to calculate profit.*

These problems might have affected retailers' acceptance decisions and lowered channel efficiency in the replication study, especially for subjects in the QD condition:

1. The literature has shown that negative emotions affect subjects' decisions of whether to accept a contract offer (e.g., Pillutla and Murnighan 1996).

2. Proposition 1 of the original paper shows that channel efficiency decreases with the complexity of the decision task (page 694). If subjects found the experimental task more complex in the QD condition, this alone could have lowered the QD's condition's efficiency in the replication study.

It is worth mentioning that participants in the original study neither had problems with the calculations nor reported frustration. They understood and completed the task within 1.5 hours in both the TPT and QD conditions.

To mitigate frustration, the replication team implemented an unusual and unplanned protocol deviation in 75% of the experimental sessions. Specifically, the team used built-in software to correct calculation mistakes made by participants. This protocol deviation might have changed the nature of the experimental task.

Finally, the replication study was performed at one location although the plan was to include two separate locations. At this one replication location, the actual number of subjects did not meet the pre-registered goal, reducing the power of the replication study from the pre-registered target of 90% to around 80%.

Conclusion

The main results of the original study were replicated. The replication study confirmed (1) that TPT is inefficient and (2) that QD helps the manufacturer and retailer design more efficient contracts compared with TPT.

Contrary to the original paper, the replication study found a lower acceptance rate in the QD condition, which made the overall efficiency statistically indistinguishable between the TPT and QD conditions. However, acceptance rate was not the primary focus in our original study. Its result in the replication study was less stable than the other channel decisions. Moreover, participants' unanticipated frustration in the QD condition alone might have affected their acceptance rate and lowered channel efficiency. The unplanned protocol deviation to correct participants' calculations could also have altered the nature of the experimental task.

Future research should seek to understand why participants in the replication found it difficult to complete the task in the QD condition even though QD is economically equivalent to TPT. Could the reason be the composition of the subject pool? Was QD computationally more challenging than TPT and, if so, how would the computational challenge differ across subject pools? It would also be interesting to examine whether a better protocol deviation, such as several rounds of practice, would improve participants' understanding of the experimental task and alleviate their frustration.

References

Ho, T.-H., Zhang, J. 2008. Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science*, 54(4), 686-700.

Pillutla, M.M., Murnighan, J.K. 1996. Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, 68(3), 208-224.

Tirole, J. 1988. *The Theory of Industrial Organization*. MIT Press, Cambridge, MA.

Author Response to Replication of Ho and Zhang (2008)
Addendum

Teck-Hua Ho and Juanjuan Zhang

February 23, 2023

We thank the replication team for informing us of their second wave of replication, conducted after our original response (dated June 11, 2022, above). We received their revised replication report on February 14, 2023.

In this second wave, the replication team collected additional data at the original replication site (at the University of Texas at Dallas) and added a secondary replication site (at the University of Michigan). We appreciate not only the effort but also the improved quality of data collection in this second wave of replication – there were no reports of subject frustration or confusion about the task, unlike in the first wave of replication.

Unfortunately, at the Michigan site, the target number of observations (N = 1,348) was not met, reducing the statistical power of the study from the pre-registered target of 90% to 60%. With this reduced sample at Michigan, efficiency was higher in the quantity discount (QD) condition than in the two-part tariff (TPT) condition (p = 0.111), which was in the same direction as our original study.

Would the efficiency result be statistically significant had the Michigan site met the target number of observations? We tested this hypothesis by pooling the data collected in the second wave of replication (396 from UT Dallas and 629 from Michigan). Indeed, efficiency is higher in the QD condition than in the TPT condition at p = 0.039, replicating our original finding (see the table below).

	Original Study			**Replication Study (Second Wave)**		
	TPT	QD		TPT	QD	
Variable	N = 264	N = 242	p-value	N = 508	N = 517	p-value
Wholesale price	3.96 (1.17)	3.41 (1.25)	0.000	4.56 (1.42)	4.32 (1.48)	0.009
Fixed fee	5.24 (2.32)	6.95 (4.17)	0.000	4.17 (2.55)	5.18 (2.75)	0.000
Retail price (if accept)	6.86 (0.54)	6.71 (0.80)	0.029	7.16 (1.01)	6.98 (0.92)	0.007
Acceptance (%)	74.24 (43.81)	82.23 (38.30)	0.029	78.15 (41.36)	80.27 (39.83)	0.403
Efficiency (%)	69.51 (41.27)	76.37 (36.18)	0.047	66.33 (39.59)	71.25 (36.77)	0.039

Notes: Values in parentheses are standard deviations.

The replication team chose to pool data by site, combining the data obtained in both waves at UT Dallas. This is problematic since the data collected at UT Dallas in the first and second waves were not comparable (due to subject frustration and confusion in the first wave).

In the second wave of replication, once again the contract terms (i.e., wholesale prices, fixed fees, retail prices) were significantly more efficient in the QD condition than in the TPT condition. It appears that these results are very robust regardless of subject heterogeneity, frustration, and confusion during the study. These were the main results of the original study, as highlighted in our original response (dated June 11, 2022, above).

We remain concerned about the continued use of the unplanned protocol deviation to correct subjects' calculation mistakes. This is problematic from a replication perspective. Hence, the conclusions drawn should be treated with caution.

Moreover, it may be interesting for the replication team to analyze the degree of calculation mistakes. This can be a good proxy for subject understanding of the experimental task. The level of understanding of the task remains a limitation to the generalizability of the replication results.

In summary, with higher-quality data collection, the second wave of replication study replicated our original findings on both the efficiency outcome and its underlying mechanism. These findings suggest that subject frustration and confusion can affect replication results.

Reply to Author Response for "Designing Pricing Contracts for Boundedly Rational Customers: Does the Framing of the Fixed Fee Matter?"

By: Teck-Hua Ho and Juanjuan Zhang

Primary Team: Elena Katok and Kyle Hyndman
University of Texas at Dallas

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

Tertiary Team: Samantha Keppler and Stephen Leider
University of Michigan

On February 23, 2023, we received the final author response about our replication efforts for Ho and Zhang (2008). This followed a series of email exchanges between the replication team and the original authors. We provide further context below but copy one important extract from their response that we believe deserves further discussion and a formal reply. Their response stated (italics added):

> *Would the efficiency result be statistically significant had the Michigan site met the target number of observations? We tested this hypothesis by pooling the data collected in the second wave of replication (396 from UT Dallas and 629 from Michigan). Indeed, efficiency is higher in the QD condition than in the TPT condition at* $p = 0.039$*, replicating our original finding (see the table below).*

	Original Study			Replication Study (Second Wave)		
	TPT	QD		TPT	QD	
Variable	*N* = 264	*N* = 242	*p*-value	*N* = 508	*N* = 517	*p*-value
Wholesale price	3.96 (1.17)	3.41 (1.25)	0.000	4.56 (1.42)	4.32 (1.48)	0.009
Fixed fee	5.24 (2.32)	6.95 (4.17)	0.000	4.17 (2.55)	5.18 (2.75)	0.000
Retail price (if accept)	6.86 (0.54)	6.71 (0.80)	0.029	7.16 (1.01)	6.98 (0.92)	0.007
Acceptance (%)	74.24 (43.81)	82.23 (38.30)	0.029	78.15 (41.36)	80.27 (39.83)	0.403
Efficiency (%)	69.51 (41.27)	76.37 (36.18)	0.047	66.33 (39.59)	71.25 (36.77)	0.039

Notes: Values in parentheses are standard deviations.

> *The replication team chose to pool data by site, combining the data obtained in both waves at UT Dallas. This is problematic since the data collected at UT Dallas in the first and second waves were not comparable (due to subject frustration and confusion in the first wave).*
>
> *In the second wave of replication, once again the contract terms (i.e., wholesale prices, fixed fees, retail prices) were significantly more efficient in the QD condition than in the TPT condition. It appears that these results are very robust regardless of subject heterogeneity, frustration, and confusion during the study. These were the*

main results of the original study, as highlighted in our original response (dated June 11, 2022, above).

Having seen an earlier version of their response, which included much the same language, on February 19, 2023 the replication team sent the original authors an email in which we sought clarification on how they defined the "second wave" in their response and also stressed to the authors that the sessions were conducted as described in Table 1. In so communicating, we wanted to emphasize that the majority of sessions conducted prior to the first submission were methodologically the same as all sessions conducted after the initial submission because they used the same experimental protocol. This is because their concerns center around the protocol changes introduced after the issues observed in the initial four sessions. We concluded our email by stating (italics added), "*If based on this clarification you would like to change how you wish to pool data across UTD and Michigan, feel free to edit your addendum.*"

On February 23, 2023, the authors responded that (italics added), "*Yes, the definition of waves is what we had in mind: the second wave refers to all data collection that happened since our initial response on June 11, 2022.*" The authors declined to change how they pool the data but they did edit their response in other dimensions, which is what we have included above. Therefore, in the "pooled analysis" in their response they do not include 12 sessions which employed the same experimental protocol as the 16 included in their analysis.

Table 1 Details on Sessions

Location	Timing	Protocol	Sessions	Note
UT Dallas	S2022	Initial	4	Prior to Original Submission
UT Dallas	S2022	Modified/Informative	12	
UT Dallas	F2022	Modified/Informative	6	After Original Submission
Michigan	F2022	Modified/Informative	8	
Michigan	S2023	Modified/Informative	2	

Notes: (1) S2022 denotes Spring 2022, F2022 denotes Fall 2022 and S2023 denotes Spring 2023. (2) Recall that the "initial" protocol required subjects to correctly input relevant numbers without any support before proceeding, while the "modified/informative" protocol still required subjects to correctly enter relevant numbers but gave subjects informative support to assist them in the event of errors.

We believe it is important to clarify a few points. First, while we did fall short of the target pre-registered sample size at one of the sites, we did not "choose to pool the data by site" in an ex post fashion. We analyzed data by site in all replications that we conducted, which also follows our pre-registration. Second, we would like to emphasize only two sessions for each treatment were conducted under the initial protocol in which subjects expressed some frustration.

In Table 2, we report the results including all sessions from the modified/informative protocol. In these results, the overall efficiency is directionally higher in the QD treatment, but the difference is not statistically significant ($p = 0.121$).

If the authors' pooling analysis is intended to speak to whether further data collection to reach greater power would likely lead to replicating the original result, these results using all the data under the modified protocol answer that question more directly. When pooled in this manner, the overall effect is directionally consistent with the original result,

Table 2 Replication Results on Efficiency (Pooling All Sessions With Modified/Informative Protocol)

Parameter	TPT		QD		p-value
Efficiency	64.84%	(40.92)	67.79%	(39.93)	0.121
Wholesale Prices	4.59	(1.49)	4.329	(1.56)	0.000
Fixed Fees	4.33	(2.63)	5.38	(2.80)	0.000
Acceptance Rate	76.66		76.16		0.804
N	904		902		
Conditional Efficiency	84.58%	(22.64)	89.00%	(14.26)	0.000
Retail Prices	7.11	(1.11)	6.90	(0.98)	0.000
N	693		687		

although not statistically significant at the required level. However, we would also like to stress that the mechanism appears to be different than in the original paper. Ho and Zhang (2008) found that conditional efficiency was not significantly different across treatments but rejections were higher in the TPT treatment. We find that there is no difference in rejections across treatments, while the conditional efficiency is significantly higher in the QD treatment.

We recognize that our replication effort did not provide an exact replication because of differences in the study designs (original: paper and pencil, information entered not checked vs replication: computerized, information entered checked). We hope the reader and the authors understand our reasons for these deviations. This was not a frivolous decision. We wished to conduct a replication over the computer to make sure that our replication itself is easily replicable. Regarding checking calculations, if subjects were allowed to enter incorrect calculations, would they have been paid based on what they entered or what was correct? If we were to pay subjects based on what they entered, this would have created an obvious incentive compatibility problem. If we were to pay them based on correct calculations, this would have created a loss of control in the experiments.

In general, while the result for this paper (in the context of the replication project) is a failure to replicate the original efficiency result, we believe there remain several open questions around what aspects of the original results and mechanisms are robust (and under what conditions), and we would encourage further study to gain greater clarity.

References

Ho, Teck-Hua, Juanjuan Zhang. 2008. Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science* **54**(4) 686–700.

F.3. Shunko et al. (2018)

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).

G. Links to Pre-Registration Details

- Chen et al. (2013): `https://aspredicted.org/5jq65.pdf`.
- Croson and Donohue (2006): `https://aspredicted.org/2y3ag.pdf`.
- Davis et al. (2011): `https://aspredicted.org/w8cp6.pdf`.
- Engelbrecht-Wiggans and Katok (2008): `https://aspredicted.org/926k7.pdf`.
- Ho and Zhang (2008): `https://aspredicted.org/ya2gu.pdf`.
- Kremer and Debo (2016): `https://aspredicted.org/b43k3.pdf`.
- Kremer et al. (2011): `https://aspredicted.org/7r4i2.pdf`.
- Özer et al. (2011): `https://aspredicted.org/id2sz.pdf`.
- Schweitzer and Cachon (2000): `https://aspredicted.org/759k8.pdf`.
- Shunko et al. (2018): `https://aspredicted.org/yq3sk.pdf`.