



Management Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

A Replication Study of Operations Management Experiments in Management Science

Andrew M. Davis, Blair Flicker, Kyle Hyndman, Elena Katok, Samantha Keppler, Stephen Leider, Xiaoyang Long, Jordan D. Tong

To cite this article:

Andrew M. Davis, Blair Flicker, Kyle Hyndman, Elena Katok, Samantha Keppler, Stephen Leider, Xiaoyang Long, Jordan D. Tong (2023) A Replication Study of Operations Management Experiments in Management Science. Management Science 69(9):4977-4991. <https://doi.org/10.1287/mnsc.2023.4866>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2023, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.









For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

A Replication Study of Operations Management Experiments in *Management Science*

Andrew M. Davis,^{a,*} Blair Flicker,^b Kyle Hyndman,^c Elena Katok,^c Samantha Keppler,^d Stephen Leider,^d Xiaoyang Long,^e Jordan D. Tong^e

^aSamuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York 14853; ^bDarla Moore School of Business, University of South Carolina, Columbia, South Carolina 29208; ^cNaveen Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080; ^dRoss School of Business, University of Michigan, Ann Arbor, Michigan 48109; ^eWisconsin School of Business, University of Wisconsin–Madison, Madison, Wisconsin 53706

*Corresponding author

Contact: adavis@cornell.edu,  <https://orcid.org/0000-0002-1689-2299> (AMD); blair.flicker@moore.sc.edu,  <https://orcid.org/0000-0003-0294-3400> (BF); kyleb.hyndman@utdallas.edu,  <https://orcid.org/0000-0003-3666-8734> (KH); ekatok@utdallas.edu,  <https://orcid.org/0000-0002-7037-7896> (EK); srmeyer@umich.edu,  <https://orcid.org/0000-0003-3329-5629> (SK); leider@umich.edu,  <https://orcid.org/0000-0002-1106-7038> (SL); xiaoyang.long@wisc.edu,  <https://orcid.org/0000-0002-3908-6441> (XL); jordan.tong@wisc.edu,  <https://orcid.org/0000-0002-1962-058X> (JDT)

Received: May 3, 2022

Revised: February 24, 2023

Accepted: April 12, 2023

Published Online in Articles in Advance:
July 11, 2023

<https://doi.org/10.1287/mnsc.2023.4866>

Copyright: © 2023 INFORMS

Abstract. Over the last two decades, researchers in operations management have increasingly leveraged laboratory experiments to identify key behavioral insights. These experiments inform behavioral theories of operations management, impacting domains including inventory, supply chain management, queuing, forecasting, and sourcing. Yet, until now, the replicability of most behavioral insights from these laboratory experiments has been untested. We remedy this with the first large-scale replication study in operations management. With the input of the wider operations management community, we identify 10 prominent experimental operations management papers published in *Management Science*, which span a variety of domains, to be the focus of our replication effort. For each paper, we conduct a high-powered replication study of the main results across multiple locations using original materials (when available and suitable). In addition, our study tests replicability in multiple modalities (in-person and online) due to laboratory closures during the COVID-19 pandemic. Our replication study contributes new knowledge about the robustness of several key behavioral theories in operations management and contributes more broadly to efforts in the operations management field to improve research transparency and reliability.

History: Accepted by David Simchi-Levi, operations management.

Funding: The authors gratefully acknowledge financial support from Cornell University, the University of South Carolina, the University of Texas at Dallas, the University of Michigan, and the University of Wisconsin–Madison.

Supplemental Material: The data files and online appendices are available at <https://doi.org/10.1287/mnsc.2023.4866>.

Keywords: replication • operations management • experiments • management science replication project

1. Introduction

In this paper, we conduct a large-scale replication study of laboratory experiments in the area of operations management published in *Management Science* prior to 2020. We focus on 10 papers that include experiments with standard laboratory participant populations, such as university students and Mechanical Turk (mTurk) workers, across five important operations management domains: inventory, supply chains, queuing/production, forecasting, and sourcing.

Our study follows a recent trend within social sciences and natural sciences to conduct large-scale high-power replications of existing empirical findings. Some notable examples of such projects include the reproducibility

project: psychology (RPP) in psychology, the experimental economics replication project (EERP) in economics, and the social sciences replication project (SSRP) in social sciences (Open Science Collaboration 2015; Camerer et al. 2016, 2018). These previous replication projects aimed to address concerns raised by the academic community around the validity of existing empirical results (Leamer 1983, Roth 1994).

There is currently a movement in many disciplines to increase the transparency and reliability of published results. Such efforts include data disclosure policies, which require authors to share original data and analysis files when a paper is published, and preregistration of research designs. In particular, preregistration requires

that, before data collection occurs, the authors specify their hypotheses, experimental design, data collection procedures, target sample size, and proposed analysis methods. Although data disclosure policies and preregistration encourage transparent and rigorous practices in future empirical research, they do not address questions raised around past results. For example, they cannot address whether a reported result is a false positive (or false negative). High-powered direct replications can help to address such concerns.

With this motivation, we conduct a large-scale replication study for experimental operations management papers published in *Management Science*. Although many related replication projects are interested in identifying a single “percent replication” statistic in a specific field (e.g., 36% for RPP, 61% for EERP, and 62% for SSRP), our primary goal is different. We aim to leverage our replication efforts to provide theoretical and practical insights to improve rigor in the field of behavioral operations. Part of this involves high-powered replication across multiple locations using original materials (when available and suitable) to further solidify, or complicate, key findings of impactful experimental papers in the operations management literature. Identifying behavioral results that robustly replicate will give researchers greater confidence in applying and extending these ideas in new areas. At the same time, published results that do not replicate or whose results are complicated by our replication findings still generate important insights about whether and how to apply and extend the published results. Furthermore, our replication also unearths new best practices for researchers in the operations management field.

By soliciting input from the broader operations management community, we identify 10 prominent experimental papers that the community expressed interest in seeing included in our replication project. Consistent with related replication studies in psychology and economics, we strive to collect sample sizes that have a statistical power of 90% to detect the original result at the 5% level. However, unlike many past studies, we go a step further and attempt to conduct two independent replications for each paper.¹ Our project, which took place from the beginning of 2020 to the beginning of 2023, involved eight researchers from five universities and included 2,514 participants.

There are multiple ways in which one can conclude that an existing result replicated or not. In our study, we base our conclusions on whether we observe a statistically significant effect in the same direction as the original study with $p < 0.05$. Other popular approaches are based on effect sizes. We report estimated effect sizes as well, but our surveys and main categorizations are based on p values. Using this approach, for any paper where the primary hypothesis was replicated at both the replication sites, we categorize this as “full”

replication. For any paper where the primary hypothesis was replicated at one of the two replication sites, we categorize this as “partial” replication. For any paper where the primary hypothesis failed to replicate at either of the two replication sites, we categorize this as “no” replication.

As mentioned previously, our goal is not to distill our entire project into a single replication statistic for the operations management field. Even papers that did not fully replicate have value in providing intriguing preliminary evidence regarding results that the research community deemed interesting enough to warrant a high-powered replication. Nevertheless, using the previously defined categorization methodology, we find that, of the 10 papers whose main hypotheses are tested in the project, 6 achieve full replication, 2 achieve partial replication, and 2 do not replicate.²

In addition to replicating existing results from well-known laboratory experiments in operations management, we also conducted a survey that solicited researchers’ predictions about which results may or may not replicate. In particular, prior to any analysis, we asked operations management researchers to predict (i) the likelihood of a result replicating and (ii) the confidence in their answer. Our analysis of these data indicate that, although the predictions were associated with the replicability of the papers considered, the replication results added information beyond the community’s prevailing beliefs. The behavioral operations management (BOM) community was generally more optimistic about the replicability of most papers; however, the BOM and non-BOM respondents had similar overall prediction accuracy.

The rest of the paper is outlined as follows: in Section 2, we provide details around our methodology and protocols, in Section 3, we outline a summary of results pertaining to both replication and the prediction survey, and in Section 4, we conclude with a short discussion. This paper provides only a summary of our work. Interested readers can find the individual replication reports, preregistrations, data and analysis files, responses from original authors, and more, on the Management Science Replication Project (MSRP) website, located at www.msreplication.com.

2. Methods

In December 2019, David Simchi-Levi, editor-in-chief of *Management Science*, issued the following challenge to the community in the “From the Editor” blog:

“The editorial board would like to publish a paper, likely a Fast Track paper, that reports replicability of laboratory experiments published by *Management Science*. This was done in economics, see Camerer et al. (2016), and in social science, see Camerer et al. (2018), and it is time to do the same for *Management Science* papers.” (Simchi-Levi 2019)

Shortly after this request, we formed a project team of eight scholars from five universities with dedicated behavioral laboratories and established participant pools. After forming the team, our first goal was to establish high-level objectives for the project. These include the following: (i) attempt to replicate between eight and ten experimental papers published in *Management Science*, (ii) target a minimum 90% power level and 5% detection level for all replications, and (iii) conduct a replication for each selected paper at two independent sites.

After setting these objectives, we turned to the operations management community to determine which papers to include in our replication effort. The replication team first identified a list of papers that report on a laboratory experiment and are published in *Management Science*, prior to the start of this project in 2020, across five core areas: inventory, supply chains, queuing/production, forecasting, and sourcing. The choice of these five areas was primarily based on the relevant chapters from *The Handbook of Behavioral Operations Management* (Donohue et al. 2019) and a list of articles created by Karen Donohue for her survey papers in behavioral operations (Donohue and Schultz 2019, Donohue et al. 2020). To narrow the list, we considered three factors: the importance of the key result, the citation count and rate, and the feasibility of replication. Regarding this last point, we chose to focus on papers that collected data from traditional experimental participant pools: university students and mTurk workers (as opposed to, for example, experienced managers).

Based on the previous criteria, we identified 24 candidate papers: 5 in inventory, 5 in supply chains, 4 in queuing/production, 5 in forecasting, and 5 in sourcing. We then designed a survey to send to the broader operations management community, asking researchers to vote on which papers they would like to see replicated in the final set of 8–10 papers. Each paper was listed with the title, year, authors, and a brief summary of the result to be replicated. Respondents could vote for up to

two papers within each category. We sent this survey to the Manufacturing and Service Operations Management Society (MSOM) and behavioral operations management sections of INFORMS and the Productions and Operations Management Society on September 4, 2020. We also sent a reminder on September 29, 2020, and later closed the survey on October 10, 2020. In total, we received 97 responses. The replication project authors did not participate in the survey. For more details relating to this survey, including the list of 24 papers and brief summaries, please see Online Appendix A.

We used the outcome of this survey to identify the final list of 10 papers to include in our replication study. Specifically, we included the two papers from each category with the highest number of votes. We then proceeded to assign each paper to two replication sites with associated replication teams (scholars from those sites). Table 1 includes a summary of the papers and other relevant details. Each paper included a primary replication team (listed first in Sites column), consisting of two team members, and a secondary replication team (listed second), also consisting of two members. In reviewing this table, one may note that some of the papers are coauthored by some of the members of the replication team. In an effort to avoid conflicts of interest, no replication member was responsible for conducting a replication of their original study.

For each paper in Table 1, we aimed to replicate a key finding with 90% statistical power at each of the two sites. To determine the target sample size for each site, we followed the power formula for a z test used in other replication studies (Camerer et al. 2016). Specifically, the required sample size fraction of the original sample size is calculated by $(3.242/z)^2$, where the z value comes from the original study. Importantly, this equation is based on detecting an effect at the 5% level (and 90% power). For those papers that did not originally conduct a z test, we converted the relevant p value into a z value and then applied the same calculation. Although

Table 1. Selected Papers and Replication Plan Details

Paper	Category	Sites	Decision type	Original N	Target N
Chen et al. (2013)	Inventory	UM and UTD	Individual	50	50
Croson and Donohue (2006)	Supply chain	UTD and UW	Group	88	254
Davis et al. (2011)	Sourcing	UW and UM	Individual	20	40
Engelbrecht-Wiggans and Katok (2008)	Sourcing	UW and CU	Individual	40	64
Ho and Zhang (2008)	Supply chain	UTD and UW	Group	94	252
Kremer and Debo (2016)	Queuing	USC and UM	Group	100	100
Kremer et al. (2011)	Forecasting	UTD and UM	Individual	86	86
Özer et al. (2011)	Forecasting	UM and CU	Group	8	40
Schweitzer and Cachon (2000)	Inventory	CU and UTD	Individual	33	40
Shunko et al. (2018)	Queuing	UW and USC	Individual	113	244
Total number of participants (N)				632	1,170

Notes. CU, Cornell University; UM, University of Michigan; USC, University of South Carolina; UTD, University of Texas at Dallas; UW, University of Wisconsin. CU and USC each had only one team member; these two individuals worked as a team. Because we recruit M-Turk subjects for Shunko et al. (2018), both “sites” used the same subject pool. N is the number of participants in the treatments of interest in the original paper. “Target N ” is the target number of participants *per site* based on the three sample-size criteria (total N of 2,340 across both sites).

a better approximation methodology is often feasible for a specific individual paper, following this method allows for a consistent procedure.

For some papers, we found that the required sample size to achieve 90% power was relatively small (e.g., less than 10 participants). Therefore, we decided to impose three criteria in determining the target sample size for each replication, at each site: (i) the target sample size must achieve a minimum power level of 90%, as outlined previously; (ii) the target sample size must be at least as large as the original study's sample size; and (iii) the target sample size must be at least 40 participants. Based on these criteria, Table 1 shows the resulting target sample sizes for each replication at each site.

Next, each team began coordinating the replication study for their assigned papers. Each primary replication team corresponded with the original authors of the selected papers. They notified the original authors that their paper had been selected for the replication project, proposed the main hypothesis to replicate, requested any original experimental materials, and solicited any other comments or suggestions from the original authors. A sample of this correspondence is provided in Online Appendix C.

After consulting with the original authors, the primary replication teams drafted a preliminary replication report for each of their assigned papers. To summarize, each preliminary report consisted of the following sections:

the hypothesis to be replicated, the power analysis, the sample details, the materials used, the experimental procedure, and the planned differences relative to the original study. At this time, there were also three sections that were left unfinished in each report, because no data had been collected (replication results, unplanned protocol deviations, and discussion).

While developing the preliminary replication reports, we completed and submitted a preregistration for each paper via www.aspredicted.org. Each preregistration included the relevant hypothesis, sample size requirements, data collection protocols, exclusion requirements, and planned data analyses (we provide direct links for each preregistration in Online Appendix G).

After the preliminary reports and preregistrations were drafted, we sent these materials to the original authors and solicited further feedback (please see Online Appendix D for an example of this correspondence). We also submitted and received approval through each site's institutional review board (IRB). Table 2 provides a list of the 10 papers and the relevant hypotheses.

Because of the COVID-19 pandemic, experimental laboratories were shut down at all universities involved in the project for several months. Therefore, the research team needed to devise an innovative replication approach different than previous replication efforts. With the goal of completing the data collections in a timely manner, the team decided that for any selected paper

Table 2. Selected Papers and Hypothesis Details

Paper	Hypothesis
Chen et al. (2013)	Subjects order higher quantities under the O-payment scheme ($-c$ per unit ordered, $+p$ per unit sold) than under the C-payment scheme ($+(p - c)$ per unit ordered, $-p$ per unit leftover), even though the two are mathematically equivalent.
Croson and Donohue (2006)	Hypothesis 3. Sharing dynamic inventory information across the supply chain will decrease the level of order oscillation.
Davis et al. (2011)	In second-price sealed bid auctions, the seller chooses a higher reserve price when the number of bidders is larger (contrary to standard theory).
Engelbrecht-Wiggans and Katok (2008)	Corresponding to hypothesis 1 (Effect of Winner's Regret), providing both "Loser's Regret" and "Winner's Regret" feedback leads to lower average bids than only providing "Loser's Regret" feedback.
Ho and Zhang (2008)	Supply chain efficiency is higher when a two-part tariff is framed as a quantity discount as opposed to a fixed fee.
Kremer and Debo (2016)	Relative to the setting with no informed consumers ($q = 0$), the presence of informed consumers ($q = 0.50$) makes uninformed consumers (a) less likely to purchase upon observing a short wait ($w = 1$) and (b) less sensitive to the purchase probability reduction associated with each marginal unit of wait time. We test these two findings in the setting with a high prior of quality ($p_0 = 0.50$, treatments Q_{00} and Q_{50}).
Kremer et al. (2011)	Hypothesis 1 (system neglect). Individuals show relatively more overreaction for low values of $W = c^2/n^2$, and relatively more underreaction for high values of W .
Özer et al. (2011)	In the $C_H U_H$ treatment, manufacturers' messages will be positively correlated with their private forecast, and suppliers' capacity decisions will be positively correlated with the messages received.
Schweitzer and Cachon (2000)	Newsvendor order quantities are set too low for high-profit products and too high for low-profit products.
Shunko et al. (2018)	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).

where the decision task involved an individual decision (e.g., an individual newsvendor decision versus a supply chain contracting interaction), we would first conduct an online-asynchronous replication using the participant pool associated with the assigned replication sites. Any papers requiring group decisions or other types of interactions among participants were to be conducted synchronously and in-person once laboratories reopened.³ A typical asynchronous replication proceeded as follows: a participant from the relevant participant pool (e.g., a university student signed up to receive notifications from a particular experimental laboratory) would sign up for a study, receive a link to the game, then read the instructions and complete the game, on their own, within a required time frame. If the hypothesis replicated in this environment, then no further data collection was required. However, if the hypothesis did not replicate for a particular site, an additional in-person data collection would take place for that site.⁴ This meant that a single paper could potentially involve four replications in total: two sites asynchronously and, if the hypothesis failed to replicate in both asynchronous collections, two sites in-person. The advantages of this alternative approach were as follows: (i) it allowed the replication project to progress despite laboratory closures due to COVID-19⁵; (ii) it created additional knowledge about the robustness of certain in-person studies to online, asynchronous conditions; and (iii) it ensured that each paper had the opportunity to be replicated in the modality in which it was originally conducted (as was standard in prior replication efforts).

The decision to treat differently individual vs. group decision tasks was made out of necessity. Most university laboratories did not have experience executing interactive group studies online. There are also technical challenges: a failed Internet connection with just one participant could invalidate an entire group. However, we acknowledge that this approach also creates differences between the replication procedures across papers: individual decision tasks that were originally run in the laboratory effectively had two chances at replication for each site—once asynchronously online and, if necessary, once in person in the laboratory. In contrast, group decision papers and those originally run asynchronously only had a single chance at replication.

Once the preregistrations were submitted, IRB approvals were received, and the replication plans were finalized for each paper—but before any data analysis was conducted, we deployed a survey to solicit the community's predictions. The survey elicited, for each of the 10 papers, respondents' beliefs of the probability (0%–100%) that the hypotheses to be tested (outlined in Table 2) would fully replicate (i.e., replicate at both sites). We also elicited their degree of confidence in their prediction. This survey was first sent to the MSOM Society and the behavioral operations management section of INFORMS on August 11,

2021. We then sent a reminder to these same communities on September 1, 2021, and had the editor-in-chief of *Management Science* (David Simchi-Levi) send yet another reminder on September 5, 2021. We closed the survey on September 17, 2021, after receiving 43 unique completed responses. The replication project authors did not participate in the survey. For more details relating to this prediction survey, please see Online Appendix B.

After the prediction survey was closed and data collections were complete, each assigned replication team conducted the relevant analyses and wrote the remaining portions of each individual replication report (which included the replication results, unplanned protocol deviations, and discussion). They then shared each respective report, along with the corresponding data and analysis files, with the original authors and provided them with the opportunity to (i) provide any feedback, (ii) prepare a response document to be posted on the MSRP website and included in the online appendix of this paper (see Online Appendix F), and (iii) ask any questions.

3. Results

Recall that, because of the COVID-19 pandemic, for any individual task experiment that was originally conducted in-person, we first replicated it remotely and asynchronously. A lack of replication in such a different experimental environment may not necessarily indicate that the original result was not replicated, but rather it provides useful information regarding robustness (e.g., differences could be due to the different experimental setting). Therefore, if a result *did* replicate for a remote-asynchronous data collection, we conclude that it replicated. If, however, a result *did not* replicate for an asynchronous data collection, then we administered a subsequent in-person data collection for that particular site and used it to draw any final conclusions. If the result replicated under a follow-up in-person replication, we conclude that the result replicated (even if the study did not replicate in a different experimental modality).

3.1. Replication

Table 3 summarizes the results for the hypotheses tested within the 10 selected papers. The first column lists the paper and relevant hypothesis. In particular, some papers included a hypothesis that consisted of two parts and hence two statistical tests. For completeness, we break each test out as a separate row in the table. The table then shows the original paper's number of participants (N), relevant p value, and estimated effect size correlation (r). Finally, it includes each specific replication's number of participants (N), relevant p value, estimated effect size correlation (r), estimated power, and replication conclusion. We code the estimated effect size as negative if the replication result is in the opposite direction

Table 3. Original and Replication Sample Sizes, p Values, Effect Sizes (r), Power, and Conclusions

Paper/hypothesis	Data type	Site	Original paper			Replication				
			N	p value	Effect size (r)	N	p value	Effect size (r)	Power	Conclusion
Chen et al. (2013)	Asynchronous	UM	50	<0.01	0.60	50	0.08	0.25	>0.99	✗
	Asynchronous	UTD	50	<0.01	0.60	50	<0.01	0.40	>0.99	✓
	In-person	UM	50	<0.01	0.60	52	<0.01	0.35	>0.99	✓
Croson and Donohue (2006)	Synchronous/in-person	UTD	88	0.06	0.21	260	0.22	0.08	0.91	✗
	In-person	UW	88	0.06	0.21	224	0.05	0.13	0.86	✓
Davis et al. (2011)	Asynchronous	UW	20	<0.01	>0.99	41	<0.01	>0.99	>0.99	✓
	Asynchronous	UM	20	<0.01	>0.99	40	<0.01	>0.99	>0.99	✓
Engelbrecht-Wiggans and Katok (2008)	Asynchronous	UW	40	0.01	0.36	69	0.22	0.10	0.92	✗
	Asynchronous	CU	40	0.01	0.36	66	0.12	0.15	0.91	✗
	In-person	UW	40	0.01	0.36	67	0.12	0.15	0.91	✗
Ho and Zhang (2008)	In-person	CU	40	0.01	0.36	66	0.05	0.21	0.91	✓
	In-person	UTD	94	0.05	0.09	263	0.85	-0.01	0.92	✗
Ho and Zhang (2008)	In-person	UM	94	0.05	0.09	120	0.11	0.06	0.60	✗
	In-person	UM	94	0.05	0.09	120	0.11	0.06	0.60	✗
Kremer and Debo (2016)	Hyp KD_a	In-person	USC	<0.01	0.83	100	<0.01	0.77	>0.99	✓
	Hyp KD_a	In-person	UM	<0.01	0.83	104	<0.01	0.89	>0.99	✓
	Hyp KD_b	In-person	USC	<0.01	0.95	100	0.02	0.46	>0.99	✓
	Hyp KD_b	In-person	UM	<0.01	0.95	104	<0.01	0.92	>0.99	✓
Kremer et al. (2011)	Hyp Kr_a	Asynchronous	UTD	<0.01	0.93	65	<0.01	0.73	>0.99	✓
	Hyp Kr_a	Asynchronous	UM	<0.01	0.93	69	<0.01	0.85	>0.99	✓
	Hyp Kr_b	Asynchronous	UTD	<0.01	0.83	70	<0.01	0.47	>0.99	✓
	Hyp Kr_b	Asynchronous	UM	<0.01	0.83	70	<0.01	0.52	>0.99	✓
Özer et al. (2011)	Hyp Oz_a	In-person	UM	<0.01	>0.99	44	<0.01	>0.99	>0.99	✓
	Hyp Oz_a	In-person	CU	<0.01	>0.99	46	<0.01	>0.99	>0.99	✓
	Hyp Oz_b	In-person	UM	<0.01	>0.99	44	<0.01	>0.99	>0.99	✓
	Hyp Oz_b	In-person	CU	<0.01	>0.99	46	<0.01	>0.99	>0.99	✓
Schweitzer and Cachon (2000)	Hyp SC_a	Asynchronous	CU	<0.01	0.74	40	<0.01	0.53	>0.99	✓
	Hyp SC_a	Asynchronous	UTD	<0.01	0.74	40	<0.01	0.68	>0.99	✓
	Hyp SC_b	Asynchronous	CU	<0.01	0.87	40	<0.01	0.84	>0.99	✓
	Hyp SC_b	Asynchronous	UTD	<0.01	0.87	40	<0.01	0.88	>0.99	✓
Shunko et al. (2018)	mTurk	UW	113	0.03	0.21	246	0.44	0.05	0.90	✗
	mTurk	USC	113	0.03	0.21	252	0.43	-0.05	0.91	✗

Notes. Some hypotheses include multiple parts, thus requiring multiple statistical tests. The following is a snapshot of the relevant part of each multipart hypothesis (see Table 2 for more details): Hyp KD_a , with informed consumers, uninformed consumers less likely to purchase for $w = 1$; Hyp KD_b , with informed consumers, uninformed consumers' purchase reduction is less sensitive with waiting time; Hyp Kr_a , more overreaction for low values of W ; Hyp Kr_b , more underreaction for high values of W ; Hyp Oz_a , manufacturers' message are positively correlated with private forecasts; Hyp Oz_b , suppliers' capacity decisions are positively correlated with messages received; Hyp SC_a , high-profit margin condition; Hyp SC_b , low-profit margin condition. CU, Cornell University; UM, University of Michigan; USC, University of South Carolina; UTD, University of Texas at Dallas; UW, University of Wisconsin. There were two authors from each university except for CU and USC; these two worked as a team. The remaining three teams were based on university affiliation. N represents the number of participants, which may differ from the number of observations used in the statistical tests (e.g., 100 participants in Kremer and Debo (2016) constitutes 25 groups of four). For Conclusion, if a result did not replicate based on the $p < 0.05$ criterion, we use ✗. If a result did replicate, we use ✓. "mTurk" represents asynchronous workers on Amazon's Mechanical Turk platform.

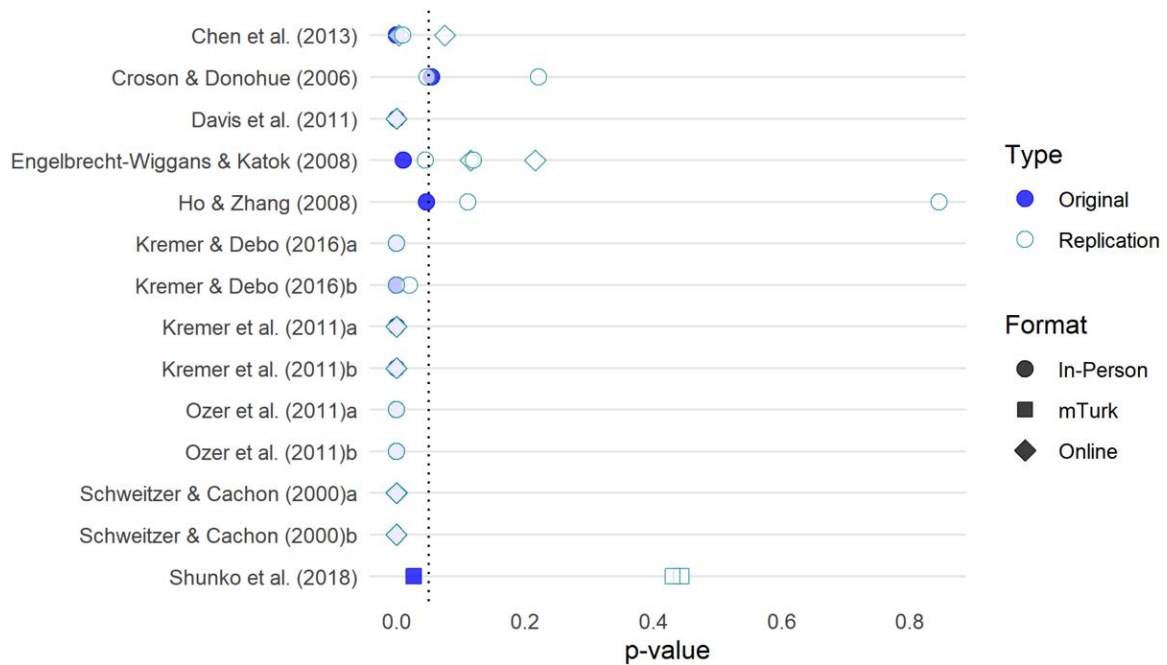
as the original study. For all power and effect size estimations, we use the same approximation strategy for all original papers and replications. Namely, as we did for determining target sample size, we use the implied z scores from the reported p values and sample sizes (irrespective of the type of test).⁶ Figures 1 and 2 depict the p values and estimated effect sizes reported in Table 3, respectively.

We now briefly discuss each paper and associated replication results (and, again, encourage interested

readers to view the full reports on www.msreplication.com):

- Chen et al. (2013) examine the effect of the structure and timing of payments on inventory decisions in a newsvendor setting, comparing "own financing" with up front costs and ex post revenue versus "customer financing" with up front profits and ex post costs for unsold items. The authors find that participants (playing the role of the newsvendor) choose higher order quantities under "own financing" than under "customer

Figure 1. (Color online) Original and Replication p Values



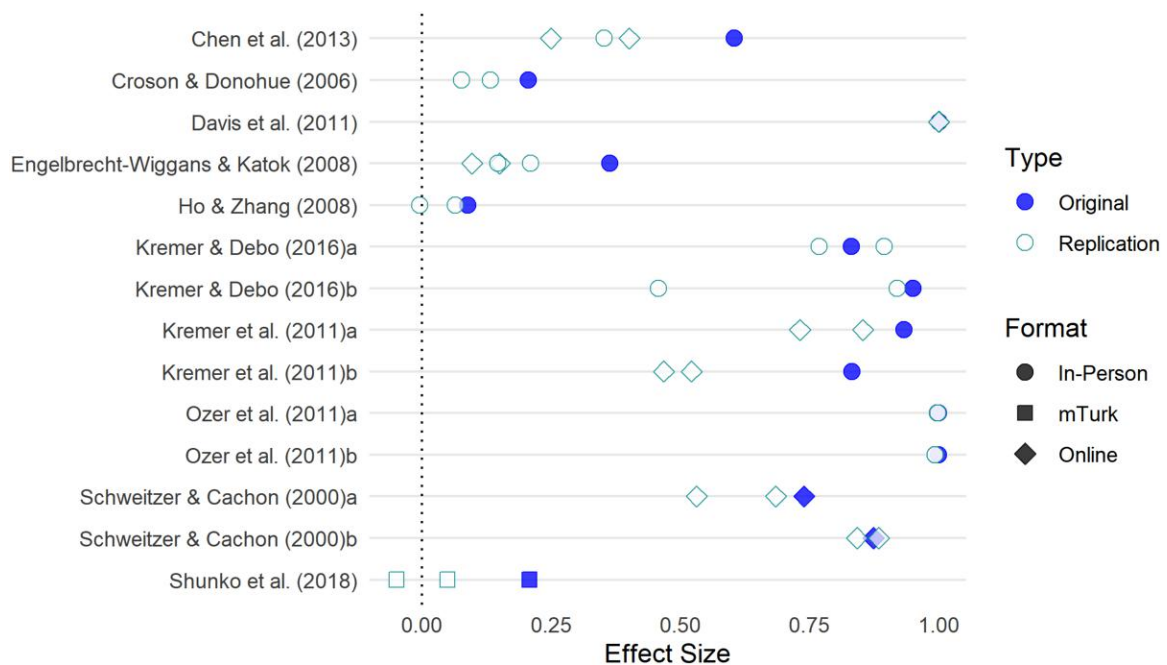
Note. Please see Table 3 notes for details around any multipart hypotheses (e.g., “a” and “b”).

financing.” We fully replicate this result. In the initial online replication wave, the results were confirmed at the $p < 0.05$ level for the secondary site, but the treatment difference had a $p = 0.0754$ at the primary replication site. However, the subsequent in-person replication at the

primary site showed a statistically significant difference with $p < 0.05$. Estimated effects sizes were somewhat smaller in the replications.

- Croson and Donohue (2006) study the bullwhip effect in a supply chain and, relevant for this replication,

Figure 2. (Color online) Original and Replication Estimated Effect Sizes



Notes. Effect sizes are coded as negative if the replication is in the opposite direction as the original study. Also, please see Table 3 notes for details around any multipart hypotheses (e.g., “a” and “b”).

whether the variance of orders is reduced when inventory information is shared. We partially replicate this result. At the primary replication site, the result is directionally consistent with Croson and Donohue (2006) but the effect is not significant, $p = 0.22$, whereas at the secondary replication site, we find a significant difference—in the same direction—as Croson and Donohue (2006) between the baseline and inventory treatment, $p = 0.048$. One observed difference that may be worth investigating is that both replication samples contained a nontrivial number of extreme outliers, whereas none were reported in the original study. However, accounting for outliers does not change the conclusions of our replication exercise. Please see Sections 3.1.1 and 3.1.2 for further discussion.

- Davis et al. (2011) investigate how auctioneers set reserve prices. Participants play the role of a seller in a second-price auction with multiple potential buyers. The authors vary the number of buyers and find that, contrary to standard theory, sellers' reserve prices are increasing in the number of buyers. We fully replicate this result in an online format (no in-person data collection was necessary). For both replications, $p < 0.001$ and the estimated effect sizes are close to one.

- Engelbrecht-Wiggans and Katok (2008) study the role of regret and feedback in bidding behavior in first-price sealed-bid auctions. They define two types of feedback. Under "Loser's Regret" participants receive feedback on the winning price and their missed opportunity to win (their resale value – winning bid). Under "Winner's regret" participants receive feedback on the second highest bid and how much money was left on the table (their bid – second highest bid). We tested their hypothesis that providing both "Winner's regret" and "Loser's regret" feedback leads to lower average bids than providing only "Loser's regret" feedback. We partially replicate this result. At both replication sites, online replications were not significant at the $p < 0.05$ level, although the results were directionally consistent with the hypothesis. Proceeding to the in-person format resulted in one out of two replications with a statistically significant finding at the $p < 0.05$ level. Estimated effect sizes in the four replications were smaller than in the original study. Please see Section 3.1.2 for further discussion.

- Ho and Zhang (2008) study whether more complex contracts can outperform simple wholesale price contracts and also whether the framing of the contract matters. They show that a quantity discount (QD) contract outperforms (i.e., leads to higher overall efficiency than) a theoretically equivalent two-part tariff (TPT) contract, and they attribute this to the fact that participants view the fixed-fee in the two-part tariff as a loss, which makes the contract less attractive and leads to significantly more rejections. We do not find that efficiency is higher in the quantity discount contract at the

primary replication site. In particular, in our replication data from the primary replication site, the overall efficiency is directionally *lower* in the quantity discount contract than the two-part tariff contract, but the difference is not statistically significant ($p = 0.845$). This is because of two countervailing results. In contrast to Ho and Zhang (2008), rejections are significantly higher in the quantity discount contract in our primary-site data. But, conditional on an agreement, the quantity discount contract is more efficient. At the secondary replication site (under-powered at 60%), we do find a directionally correct effect, with efficiency higher under the quantity discount contract. However, the effect is not significant ($p = 0.111$). Please see Sections 3.1.1 and 3.1.2 for further discussion.

- Kremer and Debo (2016) investigate purchasing decisions as a function of wait time in a setting with informed and uninformed consumers. First, they find that for short waits, the purchase probability of uninformed consumers is lower in a setting with informed consumers relative to a setting without (the "empty restaurant syndrome"). Second, as the wait time increases, uninformed consumers are more likely to purchase when there are some informed consumers in the population relative to when there are none. We find that these results fully replicate, with three $p < 0.01$ and one $p = 0.02$. We also find that three of the four effect sizes are similar to the original study with the fourth effect size being roughly half as large.

- Kremer et al. (2011) study demand forecasting behavior of participants where they systematically vary the stability of the underlying time series to be forecast. They show that participants over-react to forecast errors in stable demand environments but under-react in unstable demand environments. We show that these original results fully replicate (all four $p < 0.01$), albeit with a somewhat muted effect size, particularly in the case of high-demand variability, as participants' behavior was somewhat closer to the normative benchmark.

- Özer et al. (2011) study trust and trustworthiness in conveying market size information between a manufacturer and a supplier. The manufacturer observes a signal of the market size, and can send a cheap talk message to the supplier, who must make a capacity decision. The original authors observe substantial levels of trustworthiness (messages are positively correlated with the market signal) and trust (capacity decisions are positively correlated with message). We fully replicate these results. The correlations between message and signal, and between capacity decision and message, are statistically significant with all $p < 0.05$ and the estimated effect sizes are all close to one.

- Schweitzer and Cachon (2000) study newsvendor order decisions and find that quantities are set too low for high-profit margin products and are set too high for

low-profit margin products. Although this result has been observed in multiple newsvendor experiments, no other paper has attempted to use the design and protocols as described in their original paper (e.g., varying profit margin within subject, randomly paying just one or two participants, etc). Using their design, we find that these results fully replicate, all $p < 0.01$, with effect sizes that are similar to their original study as well.

- Shunko et al. (2018) investigate the impact of queue design on worker productivity in service systems that involve human servers. We test their hypothesis 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). Specifically, we replicate their study with mTurk workers. We are unable to replicate this result. Neither the primary nor secondary replications on mTurk result in significant differences in service times between the two queue designs at the $p < 0.05$ level. Both estimated effect sizes are less than the original, and one is negative. Although we adhere to the same mTurk selection criteria as reported in the original paper, median service times in the replication runs are larger across all conditions than those reported in the original study. Please see Section 3.1.2 for further discussion.

3.1.1. Unplanned Deviations. For 8 of the 10 papers, we were able to follow all of the methods outlined in Section 2: Two independent sites were able to collect the target sample sizes. However, there were unplanned deviations for the other two papers—Croson and Donohue (2006) and Ho and Zhang (2008)—which largely stemmed from the disruption associated with COVID-19 and the ability to recruit a sufficient number of participants. The consequence of the disruption is that for these studies, we did not achieve the target sample size at the secondary replication sites for each paper. Hence, the replication results for these two studies ought to be interpreted with these deviations in mind.

For Croson and Donohue (2006), one of the two sites (which reported a successful replication) collected data for 224 participants rather than the target sample size of 254. The estimated power for the secondary location was more than 80%, and the sample size was still far above the original study. For Ho and Zhang (2008), one site was able to exceed the target sample size of 252. Because the original secondary site was unable to collect data, a new secondary site was added. However, because of recruitment difficulties involving a combination of insufficient sign-ups and show-ups, we were unable to reach the target sample.⁷ We also made adjustments to the protocol for Ho and Zhang (2008) based on early subject feedback by making more informative the error messages that came up if a subject made a calculation error; please see the replication report for details.

3.1.2. Discussion of Papers Whose Tested Hypothesis Did Not Fully Replicate. Replication successes are relatively easy to interpret: Additional evidence lends further support to the original conclusion. However, the meaning of a replication failure is less clear. It could indicate (1) that the original finding is sound but a material difference in the replication protocol led to differential findings, (2) that a marginally significant effect sometimes crosses the significance threshold and sometimes does not, or (3) that the original effect is spurious and cannot be replicated. For each of the papers that did not fully replicate, we will briefly outline some potential explanations as informed by discussions with the original author teams.

- **Croson and Donohue (2006).** The interface of this experiment was recoded as the original software was unavailable. The original authors were consulted with this process and approved the interface. One of the key differences in the replication results was the substantially higher noise in ordering behavior across subjects at both replication locations relative to the original data. We believe that this was a contributing factor to why the experiment was only partially replicated. The question then is, what explains the starkly different degrees of noise in the original versus replication data? Although both experiments were conducted in introductory operations management courses, there was an approximately 20-year difference in when the experiments were conducted. As business (and operations management) education has changed over this period, the subject pools may differ substantially. Another possibility, which the original authors note in their response, is that the experiments at the primary replication site (UTD) were conducted in a hybrid mode (necessitated by COVID-19), with some students participating in person and others participating online. In contrast, at the secondary replication site (UW), the experiment was conducted entirely in person, as in the original experiment. Thus, it may have been more difficult to achieve common knowledge of the instructions at UTD than at UW. This could explain why the data did not replicate at UTD but did replicate at UW.

- **Engelbrecht-Wiggans and Katok (2008).** The interface of this experiment was recoded with the help of the original authors and featured minimal deviations from the original setup. Four replication studies across two sites (two online and two in-person) yielded similar results: (i) the differences in the dependent variable across conditions are in the right direction; (ii) the magnitudes of the difference are smaller in the replication studies than those in the original study; and (iii) the standard deviations are larger in the replication studies than those in the original study. More specifically, whereas the original p value is $p = 0.0103$, the p value ranges from 0.2161 to 0.3347 in the online replication studies and

from 0.0453 to 0.1196 in the in-person replication studies. These patterns are consistent with a true effect size that is smaller than that in the original study. If so, the replication experiments would have achieved lower power than that calculated based on the original p value. This could explain the nonreplication at one of the two sites.

- **Ho and Zhang (2008).** This replication was challenging even before the unanticipated deviations because the original experiment was conducted by hand while the replication was conducted using experimental software. In addition, because of a very specific matching protocol, the experiment could only be conducted in groups of 11 or 12 subjects. In the original study, participants performed all calculations manually, and these calculations were not checked. In the replication, because calculations were entered into the computer, they were checked, and participants had to enter correct calculations to proceed. This checking of the calculations, although reasonable and unavoidable in a computerized experiment, constituted a material difference that could have contributed to the difficulty of replicating the main result of this study. This was particularly true because the correctness checks require participants to spend significant time recalculating, leading to longer session times than in the original experiment (which affected recruitment and our ability to achieve the desired power). At the same time, the effect that our replication focused on—higher efficiency under the QD contract—had a high p value: $p = 0.047$. Hence, it is also possible that a marginally significant effect crossed the significance threshold in the original study (or, alternatively, failed to do so in the replication) simply due to chance.

It is also important to note that in conducting this replication project, we had to choose one result to replicate. We chose to replicate the overall efficiency result, which was agreed to by the original authors. A deeper analysis reveals, however, that some key results in the paper, which we did not attempt to replicate, but nevertheless have data on due to our replication effort, are robust. Specifically, at both replication sites, wholesale prices are significantly lower under QD than under TPT and fixed fees are higher—results consistent with the original paper. So the hypothesized behavioral mechanism, namely, the supplier's belief about the retailer's loss aversion with respect to the fixed fee, appears to be robust. What appears not to be robust, however, is the actual retailer's loss aversion with respect to the fixed fee, because the acceptance rate is not significantly higher under QD than under TPT at either of the replication sites. This higher acceptance rate, absent in the replications, was the key driver of the higher QD efficiency in the original study, which we failed to replicate.⁸

- **Shunko et al. (2018).** We chose to replicate the original mTurk experiment with an mTurk subject pool. We used the same selection criteria as the original study:

subjects were recruited from the pool of U.S.-based workers with at least 70% positive feedback and 50 successfully completed prior tasks. As in the original study, we removed subjects who did not complete the study or completed it more than once (based on duplicate IP addresses or user identification codes). During a pilot study to test the software, however, we observed a significant proportion of subjects who did not complete any cart. Given that no such data were reported in the original study, we consulted with the original authors and were informed that such data were removed from their sample; hence, we chose to do the same (as documented in the preregistration).

Nevertheless, there is evidence of differences between the two subject pools. The second-half median service times were substantially 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). As the original authors note in their response, these patterns suggest that queues were generally longer in the replications than in the original experiment, which is consequential if subjects behave differently under high (versus low) workloads. Longer service times could also indicate lower subject pool quality. The authors point out there are other observable differences in the subject pools: For nonmanipulated control variables (e.g., whether the subject is male), we observe different coefficients (see table 3 in the Shunko et al. (2018) replication report). We agree that subject pool differences is a plausible explanation for the nonreplication.

The authors shared with us partial code for the queue simulator and Qualtrics design, which we adapted to recreate the experiment as described in the original paper. In particular, we recreated the “outer shell” interface from scratch, based on the static screenshots provided in Shunko et al. (2018). The original authors tested and approved of the new interfaces (videos of the exact experiment process and the stimuli we used are available in the supplementary materials), but we cannot fully rule out the possibility that reprogramming generated meaningful differences. Another difference that we think is an unlikely explanation is that we increased the fixed rate from \$1.25 to \$3.00 to account for inflation and minimum payment standards. Finally, we note that, although this replication was conducted two times on mTurk (corresponding to the “two sites”), two mTurk runs may be more correlated than two laboratory experiments conducted at different universities.

3.2. Prediction Survey

We now examine the responses to the prediction survey and their connection to the replication results. The prediction survey elicited respondents' predictions for each paper of the likelihood (from 0% to 100%) that the results for each paper would fully replicate (i.e., obtain statistically significant results at both replication sites).

In addition, respondents were asked to indicate the confidence of their prediction on a scale of one to seven, with seven indicating the greatest confidence.⁹ We also asked respondents to report application areas in which they conduct research (including BOM), whether they have published a paper including a laboratory experiment, their familiarity with the 10 papers being replicated, and whether they are an author on one of those papers. See Online Appendix B for additional details.

In the fall of 2021, the prediction survey was sent out to the MSOM Society and the behavioral operations management section of INFORMS. We received 43 complete, unique responses, with 21 respondents indicating behavioral operations as a research area.¹⁰ Table 4 reports for each paper the average replication probability elicited from all respondents, as well as separately for BOM and non-BOM respondents. In general, we see that for all papers the average predictions were largely optimistic; however, there was a substantial range across the papers (averages from approximately 60% to 80% chance of replicating). Additionally, we see that respondents who conduct research in BOM have generally more optimistic predictions: almost all papers have more positive predictions from BOM respondents, with five papers having statistically significant differences (marginally, once we account for multiple hypothesis testing).¹¹ When we pool together all papers, we see a statistically significant overall trend, where BOM respondents' predictions are approximately 8% more positive, on average. This result is robust to a number of alternate specifications, including additional respondent controls, focusing on predictions with a confidence rating of at

least four (out of seven), and two-way clustering on both paper and respondent. Additionally, predictions that are assigned more confidence are significantly more positive (average prediction is 53.3% when confidence is one or two versus 84.6% when confidence is seven; nonparametric test for trends yields $p < 0.01$).

We also examine whether there is any association between respondents' predictions and the replication results we obtain for each paper. Comparing the average predictions for the six papers that fully replicate versus the four papers that do not, we find that the fully replicating papers have more optimistic predictions (mean of 74.3% versus 68.6%). Regressing predictions on an indicator for the paper having a full replication shows a statistically significant difference between the sets of predictions ($\beta = 5.708$, clustered standard error = 1.930, $p = 0.005$). This association is robust to additional respondent controls and restricting to high confidence predictions, as well as some specifications with two-way clustering. BOM-focused respondents have a directionally (but not statistically significant) smaller difference in predictions for papers that fully replicate or not. We can similarly look at the relationship between predictions and the relative effect sizes observed in the replications versus the original paper. To construct the relative effect sizes, we take the average of the effect sizes across each replication (as reported in Table 3) divided by the original effect size. For papers with two hypotheses, we then take the average of the two ratios. The resulting effect size ratio ranges from zero for Shunko et al. (2018) all the way to one for Davis et al. (2011) and Özer et al. (2011), where the replications find effects of the same magnitude

Table 4. Replication Predictions

Paper	Average prediction (%)			BOM vs. non-BOM		
	All	BOM	Non-BOM	Difference	<i>p</i> values	<i>q</i> values
Chen et al. (2013)	73.65	72.14	75.09	-2.95	0.942	0.394
Crosan and Donohue (2006)	71.47	78.10	65.14	12.96	0.013	0.059
Davis et al. (2011)	80.21	82.30	78.32	3.98	0.240	0.177
Engelbrecht-Wiggans and Katok (2008)	75.30	78.89	72.05	6.84	0.147	0.140
Ho and Zhang (2008)	61.81	66.05	57.77	8.28	0.192	0.160
Kremer and Debo (2016)	66.86	68.19	65.52	2.67	0.406	0.292
Kremer et al. (2011)	67.81	72.55	63.50	9.05	0.040	0.059
Özer et al. (2011)	75.07	80.05	70.10	9.95	0.022	0.059
Schweitzer and Cachon (2000)	81.95	88.24	75.95	12.29	0.021	0.059
Shunko et al. (2018)	66.21	73.90	58.86	15.04	0.019	0.059
Pooled regression estimate for BOM vs. non-BOM: $\beta = 7.83$ (standard error = 3.20, $p = 0.019$)						

Notes. This table reports the average elicited probability of “full replication” (0%–100%) for each paper, separating out respondents who self-identified as having a research focus in behavioral operations management (BOM) or not (non-BOM). The *p* values are from a rank sum test comparing the distribution of predictions for BOM and non-BOM respondents. The *q* values are “two-stage sharpened *q* values” based on the rank sum test results (following Benjamini et al. (2006) and Anderson (2008)) to account for multiple hypothesis testing. The “pooled regression estimate” is from a regression of predictions on an indicator variable for BOM focus, as well as fixed effects for each paper, and with standard errors clustered at the respondent level. The results are robust to additional controls (Respondent laboratory experience, paper familiarity and author status), to restricting to predictions with a confidence ≥ 4 , and to two-way clustered errors at the paper and respondent level.

as the original. Regressing predictions on this relative effect size ratio, we again find a positive association with a nearly 10-percentage-point increase in predictions for a paper with a relative effect size ratio of one versus zero ($\beta = 11.909$, clustered standard error = 2.659, $p < 0.001$). As before, BOM-focused respondents have a directionally smaller (but not significantly so) coefficient. Results are robust to additional respondent controls, focusing on high confidence predictions, and two-way clustering.

Finally, we compare the aggregate prediction accuracy between BOM and non-BOM respondents. For each respondent, we calculate their mean absolute prediction error (MAPE) as follows: Each paper i has a prediction error of $(100 - \text{prediction}_i)$ if it fully replicated or (prediction_i) if it did not. A respondent's MAPE is then the average over the 10 errors. We also construct a weighted MAPE where each paper is weighted by the respondent's confidence in their prediction (with higher confidence having more weight). The average MAPE for all respondents is 42.36%, whereas the Weighted MAPE is 40.45%. This reflects a small, but significant, improvement over the completely uninformative prior of a 50% replication probability (sign-rank test $p < 0.01$ for both). The average MAPE is quite similar between BOM and non-BOM respondents (42.55 versus 42.12, rank sum $p > 0.20$). The same holds for the confidence-weighted MAPE. BOM respondents had the largest increase in optimism over non-BOM respondents for two (of the four) papers that did not fully replicate, offsetting the accuracy benefits of generally being more optimistic for the other replicating papers. Taken together, these results suggest that while there is real information about

replicability captured in the predictions of the community, there is substantial information to be gained from the replication results over the prevailing sentiment in the operations management research community.

4. Discussion

This large-scale replication study has at least three important implications for the operations management community. First, it creates new knowledge about the validity, and in some cases limitations, of some of the most prominent laboratory experimental results in our field. Behavioral researchers, and those who draw on behavioral insights in analytical or other empirical work, can leverage our findings as a source of confidence about the results we test on a large scale. Second, our study contributes important insights as to the transferability of findings between the in-person and online modalities. As the COVID-19 pandemic continues to induce online modes of data collection and remote interactions with participants, it is important to have results that indicate when results can replicate across modalities and, when they do not, some discussion of why. Third, our study initiates what we hope becomes a new tradition in operations management. We believe that our study can serve as a foundation for similar operations management replication projects in the future.

4.1. Emerging Science of Replication

There are many tradeoffs and compromises to negotiate in service of replication objectives. There have now been several replication projects in psychology, economics, and operations (Table 5), and each has struck a different

Table 5. Management Science Replication Project Relative to Past Efforts

Label	Replication project			Included papers/effects				Results	
	Field	Year	Authors	Count	Multiple sites	Journals	Year(s)	Success	ES ratio
MSRP	Behavioral operations management	2023	8	10	Yes	1	2000–2018	70%	80%
RPP	Experimental psychology	2015	270	100	No	3	2008	36%	49%
EERP	Experimental economics	2016	18	18	No	2	2011–2014	61%	66%
SSRP	Experimental social science	2018	24	21	No	2	2010–2015	62%	54%
BOMT	Behavioral operations management	2018	3	3	No	2	2008	67%	n.r.
ML1	Experimental psychology	2014	51	13	Yes	12	1936–2013	77%	n.r.
ML2	Experimental psychology	2018	190	28	Yes	16	1977–2014	50%	n.r.
ML3	Experimental psychology	2016	64	10	Yes	7	1935–2013	30%	n.r.

Notes. MSRP, our project; RPP, replication project: psychology (Open Science Collaboration 2015); EERP, experimental economics replication project (Camerer et al. 2016); SSRP, social science replication project (Camerer et al. 2018); BOMT, behavioral operations experiments on mTurk (Lee et al. 2018); ML1, ML2, and ML3, Many Labs 1 (Klein et al. 2014), 2 (Klein et al. 2018), and 3 (Ebersole et al. 2016), respectively. "Multiple sites" indicates whether each included paper/effect was tested at more than one research site. "Success" indicates the proportion of replication results that are significant at the designated threshold (typically $p < 0.05$) and are in the same direction as the original result. For our project, we used the paper-site as the unit of analysis (i.e., we considered the success of each paper at each site, separately). Had we used the hypothesis-site as the unit of analysis (several fully replicated papers included two hypotheses), our replication success rate would have been 79%. "ES ratio" is the ratio of the replication effect size to the original effect size. Again, we used the paper-site as the unit of analysis. For papers that contain two hypotheses, we average the two effect sizes together before taking the ratio. Had we used the hypothesis-site as the unit of analysis, the ratio would have been 83%. n.r., not reported. RPP and SSRP do not report ES ratio explicitly—only the mean (standardized) effect size (r) for the original and replication effects. For RPP, these are 0.197 and 0.403, respectively (hence, 49%). For SSRP, they are 0.249 and 0.460, respectively (hence, 54%). For both metrics related to our project, we only included the final replication attempt at each site (i.e., we excluded the asynchronous results if a replication was repeated in-person as outlined in Section 2).

balance of competing interests. Here, we discuss some of these interrelated considerations and describe how we decided to navigate them.

4.1.1. Paper Selection Method (Mechanical vs. Curated).

When selecting which papers to replicate, some projects take a mechanical approach in which they define a small set of journals and a limited timeframe and replicate all (EERP and SSRP) or most (RPP) of the suitable papers that meet the criteria. In other projects—notably the Many Labs Projects (i.e., ML1, ML2, and ML3)—the authors curate a list of papers (often influenced by the results of an open nomination process) from a wide array of journals spanning many decades. Both approaches have merit. Mechanical inclusion sheds some light on the editorial process of particular journals and therefore may generalize to other papers published in the same journal. But inclusion rules are often arbitrary by necessity (e.g., RPP, the largest replication project to date, only includes papers published in the early months of the year 2008). Curated inclusion allows effort to be targeted to papers where replication is deemed to be most necessary or useful. However, hand-picking papers means that they are unlikely to be representative of any field or journal, so generalizability beyond the included effects may be limited.

We conducted a two-stage paper selection process meant to be a hybrid of mechanical and curated inclusion (see Section 2 for more details). The first stage of paper identification was primarily mechanical. The second stage of soliciting anonymous voting from the operations community was primarily curated. Mechanical paper identification served to help create a representative list of papers to vote on. However, by including community feedback, we hope that our replication results are more useful than had we adopted a purely mechanical approach.

4.1.2. Author Team Size and Structure (Scope vs. Standardization). All else equal, replicating more papers is better. However, each paper comes with a substantial marginal cost in money, time, effort, and coordination. Projects that attempt to replicate more papers typically have larger project teams (Table 5) and tend to operate in a decentralized fashion with an open call for collaborators (e.g., RPP, ML1, ML2, ML3). However, coordination in this setting is difficult. The Many Labs projects (especially ML1) demonstrate substantial variation in effect size from laboratory to laboratory, which could be partially attributed to variations in protocols across the different sites.

By forming a centralized eight-author team, we believe we were able to achieve a greater level of process consistency and coordination relative to a decentralized project. The five research sites are all housed in the business schools of large research universities with active, comparably outfitted behavioral labs. All project team members have previously conducted behavioral research in the

field of operations. The entire project team communicated regularly and made decisions jointly.

4.1.3. Number of Sites per Paper (Single vs. Multiple).

Another critical decision in replication projects is the number of sites for each paper. Attempting each replication at only one site, as is common among many projects, allows for a relatively large number of papers to be included in the project, but it also creates a perfect confound between paper and site. In light of this, the Many Labs Project (ML1) tested the same 13 effects at each of 36 different research sites. Their results showed that some effects were consistently more reliable than others, but also that there was substantial variation among sites. Similarly, we set out to replicate each paper at multiple sites to mitigate idiosyncratic differences among labs. Our approach turned out to be informative, as 2 of the 10 papers replicated at one site but not the other.

4.1.4. Replication Type (Exact vs. Close). Finally, replication studies may differ in the degree to which they adhere to the original experimental protocol, which in turn relates to the goal of the replication and the interpretation of the results. Chen et al. (2021) use the terms “exact replication” (identical protocol including instructions and modality) and “close replication” (some material difference in protocol which is explicitly documented) to classify replication efforts. Lee et al. (2018) provide an example of a project (labeled BOMT in figure 5) intentionally designed to be a close replication: Their goal was to test whether three selected laboratory findings would replicate on mTurk. Although we initially aimed for exact replications, several logistical challenges (e.g., laboratory closures due to COVID-19, inability to find or use original materials) resulted in close replications in some cases. The benefit of close replications over exact replications is that they can illuminate more about the robustness, or boundary conditions, of the original results.

4.2. Best Practices for Experimental Studies

The current replication project is part of a wider movement toward greater research transparency. Practices that were once rare have become standard. For example, for researchers who conduct laboratory experiments, it is now commonly expected that studies are pre-registered for documentation about the study design, hypotheses, analyses, sampling, and exclusions. It is also increasingly common that researchers make available all materials used in the laboratory, including instruction documents, associated surveys, and experimental treatment code (i.e., zTree, SoPHIE). Indeed, there is an explicit *Management Science* requirement to include data analysis code, log files, and the raw experimental data itself for laboratory experiments. The replication team can attest first-hand that such practices would indeed help improve fidelity in replication attempts and thus

more valuable knowledge can be created about replicability. Such practices would also facilitate efforts to test replicability with different participants (i.e., managers instead of students) and over time, which helps evolve behavioral theories of operations management.

Although transparency into the *implementation* of laboratory experiments is a necessary condition for evaluating replicability and robustness of the results, our experience with the replication project has illuminated that there may also be ways to improve replicability in the *design* phase of research. In particular, our research team has identified several best practices that are presently under-attended but that we believe to be critically important. For one, we recommend that for studies which aim to identify underlying psychological or behavioral truths, it is important to *design for a generic subject pool*. For example, if the experiment involves a complex mathematical calculation that is deemed trivial for the students in one university's subject pool but could be challenging in different subject pools, then adjust the design accordingly so that the result is not driven by the subject pool's attributes (i.e., provide calculators or a formula sheet, check calculations, etc.). When that is not possible, researchers may want to speculate about how their subject pool may vary from other subject pools and thus potential boundary conditions. What our own replication study, and many others, have demonstrated repeatedly is that subject pools are different, and we recommend that authors design experiments with that in mind.

Building on this, we recommend that experimental researchers increase the information they share about their design decisions, specifically distinguishing between *rigid versus flexible design decisions*. In conversations with the authors, we learned about how they saw design decisions as more or less flexible. Rigid design decisions were perceived as critical and driving the results. Meanwhile, flexible design decisions were perceived to be design decisions made for other reasons (i.e., practicality), but not theory relevant for the tested hypotheses. Moving forward, we encourage authors to include this information as part of their methods and experimental design sections. This is valuable information for any researchers building on a paper and advancing behavioral theory, whether by replicating or extending the results.

Finally, we recommend that for all researchers, it is important to *design for an online future*. Our replication project is unique in that it occurred coincident with COVID-19, where all papers included were implemented before COVID, and many had to be adapted because of one of the following: (i) the original experiment was conducted manually; (ii) the original experiment was conducted using software, but still had aspects to it that required in-person implementation; or (iii) they were interactive experiments that had to be conducted in-person, but norms of attendance and payment changed

post-COVID onset. Given the shifts that we have observed, we strongly believe that all experiments should be designed with flexibility in mind. For individual decision making experiments, this could mean designing the experiment so that it can be conducted entirely remotely. For experiments where subjects interact, a fully online implementation may not be feasible; however, rigid/inflexible designs (across various dimensions like format, number of subjects, matching protocol, etc.) have significant costs. One should think carefully about these costs and weigh them against any benefits before making any such design choice.

4.3. Limitations and Conclusion

Our replication study is not without limitations. For one, we test the replicability of only one or two key results for ten individual papers. This is relatively small compared with the results found across all published experimental operations management papers. This disproportionality was exacerbated by the fact that our study was the first large-scale replication project in our field, requiring us to consider including experimental papers published over a two-decade period. Our findings should be interpreted with this in mind. Specifically, one should take caution in using our results to make broad inferences about the replicability of other experimental studies. Furthermore, each individual paper often contains multiple results, yet we have focused only on one (or two) results. Second, replication feasibility was a necessary condition for inclusion in our replication study, meaning that we did not consider papers with field experiments or laboratory experiments with practitioner participants. This highlights an inherent tension between two concurrent movements in the operations management field: replicability and external validity. As behavioral researchers in operations management enhance external validity with field experiments or laboratory experiments involving practitioners, replicability and transparency can decrease. A final limitation to keep in mind is that the papers we include in the replication were all published in *Management Science*. Although this is a good place to start, there is more work to be done.

Acknowledgments

The authors thank seminar participants at Cornell University, Georgetown University, Syracuse University, and the Annual Behavioral Operations Management Conference for feedback; Karen Donohue for sharing a list of existing behavioral operations management experiments; and David Simchi-Levi for initiating this project.

Endnotes

¹ A series of replication projects—dubbed Many Labs 1 (Klein et al. 2014), 2 (Klein et al. 2018), and 3 (Ebersole et al. 2016)—were designed specifically to test the variability of several prominent psychological effects *between* labs. We discuss these projects further in Section 4.1.

² For one paper that achieved no replication, our collected sample size did not achieve 90% power at one of the two replication sites. See Section 3.1.1 for more details.

³ The target sample for replications for Croson and Donohue (2006) was undergraduate students taking an operations management course. Because some classes were remote/hybrid during our study, one of the two replications included a mixture of students participating remotely and students participating in-person.

⁴ The purpose of the second-stage in-person data collection for individual decision tasks was to provide a replication attempt in an environment similar to the original paper, most of which were conducted in the laboratory. One paper involving an individual decision task—Shunko et al. (2018)—contained multiple experiments, some with university students and some on mTurk. For this study, we decided to attempt replication on mTurk only without the possibility of a second-stage in-person attempt.

⁵ Laboratory closures in the wake of COVID-19 was the primary factor in our decision to replicate individual decision tasks asynchronously first. When feasible, replications should be done using the same modality as the original paper (unless they are *close replications*—see Section 4.1—explicitly designed to test robustness across different modalities).

⁶ Note the N used to estimate effect size and power is the number of observations treated as independent in the analysis, which is not necessarily the number of participants. Details and R code are available on the MSRP website.

⁷ Ho and Zhang (2008) required group sizes of 11–12 participants, which led to a number of planned sessions being canceled for lack of participants. Nevertheless, the sample size at the secondary location still exceeded that of the original study.

⁸ This discussion suggests a deeper issue in designing replication projects. There is a tradeoff between the breadth of replication across many papers and the depth of replication that can be done for any one paper.

⁹ The prediction survey is modeled after similar approaches used in other replication projects, for example, Dreber et al. (2015), DellaVigna et al. (2019), and Camerer et al. (2016).

¹⁰ Three respondents indicated that they were an author on one of the replication papers. All results are robust to excluding these responses.

¹¹ We follow Benjamini et al. (2006) and Anderson (2008) to account for multiple hypothesis tests.

References

Anderson ML (2008) Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *J. Amer. Statist. Assoc.* 103(484):1481–1495.

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rates. *Biometrika* 93(3):491–507.

Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, Kirchler M, et al. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–1436.

Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, Kirchler M, et al. (2018) Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Hum. Behav.* 2:637–644.

Chen L, Kök AG, Tong JD (2013) The effect of payment schemes on inventory decisions: The role of mental accounting. *Management Sci.* 59(2):436–451.

Chen R, Chen Y, Riyanto YE (2021) Best practices in replication: A case study of common information in coordination games. *Experiment. Econom.* 24:2–30.

Croson R, Donohue K (2006) Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Sci.* 52(3):323–336.

Davis AM, Katok E, Kwasnica AM (2011) Do auctioneers pick optimal reserve prices? *Management Sci.* 57(1):177–192.

DellaVigna S, Pope D, Vivaldi E (2019) Predict science to improve science. *Science* 366(6464):428–429.

Donohue K, Schultz K (2019) The future is bright: Recent trends and emerging topics in behavioral operations. Donohue K, Katok E, Leider S, eds. *The Handbook of Behavioral Operations Management* (Wiley, New York), 619–651.

Donohue K, Katok E, Leider S, eds. (2019) *The Handbook of Behavioral Operations* (Wiley, New York).

Donohue K, Özer Ö, Zheng Y (2020) Behavioral operations: Past, present, and future. *Manufacturing Service Oper. Management* 22(1):191–202.

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, et al. (2015) Using prediction markets to estimate the reproducibility of scientific research. *Nature* 521(7578):15343–15347.

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, Baranski E, et al. (2016) Many labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Experiment. Soc. Psych.* 67:68–82.

Engelbrecht-Wiggans R, Katok E (2008) Regret and feedback information in first-price sealed-bid auctions. *Management Sci.* 54(4):808–819.

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

Klein R, Ratliff K, Vianello M, Adams R Jr, Bahník S, Bernstein M, Bocian K, et al. (2014) Data from investigating variation in replicability: A “many labs” replication project. *J. Open Psych. Data* 2(1):e4.

Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB Jr, Alper S, Aveyard M, et al. (2018) Many labs 2: Investigating variation in replicability across samples and settings. *Adv. Methods Practice Psych. Sci.* 1(4):443–490.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Sci.* 62(10):3023–3038.

Kremer M, Moritz B, Siemsen E (2011) Demand forecasting behavior: System neglect and change detection. *Management Sci.* 57(10):1827–1843.

Leamer EE (1983) Let’s take the con out of econometrics. *Amer. Econom. Rev.* 73(1):31–43.

Lee YS, Seo YW, Siemsen E (2018) Running behavioral operations experiments using amazon’s mechanical turk. *Production Oper. Management* 27(5):973–989.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.

Özer Ö, Zheng Y, Chen K-Y (2011) Trust in forecast information sharing. *Management Sci.* 57(6):1111–1137.

Roth AE (1994) Lets keep the con out of experimental econ.: A methodological note. *Empirical Econom.* 19:279–289.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Sci.* 46(3):404–420.

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: The behavioral impact of queueing design on service time. *Management Sci.* 64(1):453–473.

Simchi-Levi D (2019) Management science: From the editor, January 2020. *Management Sci.* 66(1):1–4.