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

By: Rachel Croson and Karen Donohue

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

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

Croson and Donohue (2006) study the Bullwhip Effect in a multi-tier supply chain consisting of a Manufacturer, Distributor, Wholesaler and Retailer. In two studies, they vary whether inventory information is or is not shared and show that inventory information sharing helps to alleviate (but not eliminate) the bullwhip effect.

Hypothesis to replicate:

Hypothesis 3. Sharing dynamic inventory information across the supply chain will decrease the level of order oscillation.

Power Analysis

The original test of Hypothesis 3 is a test of the variance of orders (subject level) between the two treatments. The authors conduct a Mann-Whitney rank sum test, with 44 observations per treatment and report a z-statistic of 1.92 ($p = 0.056$).[1] Based on their statistical analysis, we estimated that to achieve 90% power, we needed 254 subjects, which corresponds to approximately 32 groups of 4 subjects for each of two treatments.

Sample

Participants for the original study were "undergraduate business students at the University of Minnesota enrolled in an introductory operations management course". The target sample size for the primary replication was 254 University of Texas at Dallas students enrolled in the class "Operations Management." The target sample size for the secondary replication was 254 subjects from the Business School at the University of Wisconsin-Madison also enrolled in the class "Operations Management."

Materials

The original authors provided us with written instructions for one of the treatments, along with a comprehension quiz. We relied on these instructions, modifying them as necessary in order to conduct the experiment using a web-based interface.

[1] Using their actual data, more precisely, the test statistic is $z = 1.915$ ($p = 0.0554$).

Procedure

We followed the same protocols outlined in sections 2.2 and 2.3 of the original paper with some minor modifications as noted later.

Analysis

The analysis is identical to the original article: a non-parametric Mann-Whitney rank sum test comparing the order variance of subjects in each of the two treatments. We also conducted additional post hoc analyses to account for some notable differences between the original data and the replication data, as we discuss later.

Differences from Original Study

1. The original experiments were run in a classroom environment at the University of Minnesota, with one class for each treatment. The two treatments were conducted back-to-back on the same day. Because our required sample size was significantly larger, we had to conduct the experiment across several different classrooms and it was not possible to achieve a "back-to-back" running of sessions at UTD. Indeed, because of the Covid-19 pandemic and issues with conducting sessions in the hybrid teaching format, sessions were conducted in both Spring 2021 and Spring 2022. At UW-Madison, all sessions were conducted on the same day (but the sample size fell slightly short of the target).

2. We used the SoPHIE online platform to run the experiment. Some students who participated in the experiment did so online from their own homes, rather than physically from the classroom. This was necessary due to the hybrid teaching format adopted by many instructors of the relevant class at UTD. All subjects at UW-Madison completed the experiment in person.

Replication Results

We collected data from 260 students at UTD and another 224 students at UW-Madison. The results are displayed in Table 1. We report the average (across subjects) order variance, the standard deviation as well as the median order variance. For ease of comparison we also report the same summary statistics for the original data. One difference from the original study that is immediately apparent is that both the UTD and UW-Madison data had a non-trivial number of extreme outliers, which severely distorts the averages. However, when looking at the medians, the results are more consistent with Croson and Donohue (2006).

In the original paper, Croson and Donohue (2006) report a Mann-Whitney rank-sum test to compare the two treatments. Under the assumption that one distribution is simply a location shift of the other distribution, then this test can be interpreted as a test of medians. However, if the location-shift assumption does not hold, then the Mann-Whitney test is best-interpreted as a test of equality of distributions (Hyndman and Embrey 2018). We will let the reader judge for him/herself about the tenability of the location-shift. Specifically, in Figure 1, we plot the empirical CDFs of the order variance for each of the three datasets. Regardless, the Mann-Whitney test, as used in the original paper, is an appropriate test given the data. Moreover, in both replication samples, the direction of the difference in distributions is the same as in the original data. For the primary replication site (UTD) the Mann-Whitney test gives $p = 0.221$, indicating that the result fails to replicate. However, for the secondary replication site (UW-Madison), the Mann-Whitney test gives $p = 0.048$, indicating that the result does replicate. Thus, the overall result is a partial replication of the hypothesis: "Hypothesis 3. Sharing dynamic inventory information across the supply chain will decrease the level of order oscillation."

Unplanned Protocol Deviations

While we had originally intended to complete data collection in the same semester, this was not feasible at UT Dallas due to Covid-19. Additionally, due to differing attendance across sessions, while the total sample of subjects who participated at UTD (260) exceeds the number for our power calculations, the samples are not balanced, with 124 subjects participating in the baseline treatment and 136 subjects participating in the inventory treatment.

At UW-Madison, all data was collected in a single semester. However, the overall sample was only 224 subjects, spread equally across the two treatments. Therefore, the desired 90% power was not achieved. We estimate that the actual power was approximately 86.5%.

Discussion

As we have noted, in both replication samples, there is a non-trivial number of apparent outliers, some of which have order variances which are orders of magnitude higher than in the original study. This was unexpected as the original data did not appear to include any such extreme outliers. We experimented with different approaches to deal with outliers, from ad hoc methods such as visual inspection to more data-driven methods such as dropping observations which are more than 3 standard deviations from the treatment average. However, none of these methods change the overall conclusion of our replication exercise — namely, that the treatment difference is in the same direction as in the original study, but the treatment difference is only statistically significant at UW-Madison, indicating a partial replication.

Table 1 Summary Results on Order Variance

	UTD			UW-Madison			Original		
	Mean	S.D.	Med.	Mean	S.D.	Med.	Mean	S.D.	Med.
Baseline	8434.06	27620	20.46	1147.99	6609	30.91	32.59	35.21	17.30
Inventory	4346.71	21671	17.58	4646.85	22490	19.83	21.67	23.04	13.22
$p-$value	0.184		0.221	0.116		0.048	0.089		0.056
Test	t		M.W.	t		M.W.	t		M.W.

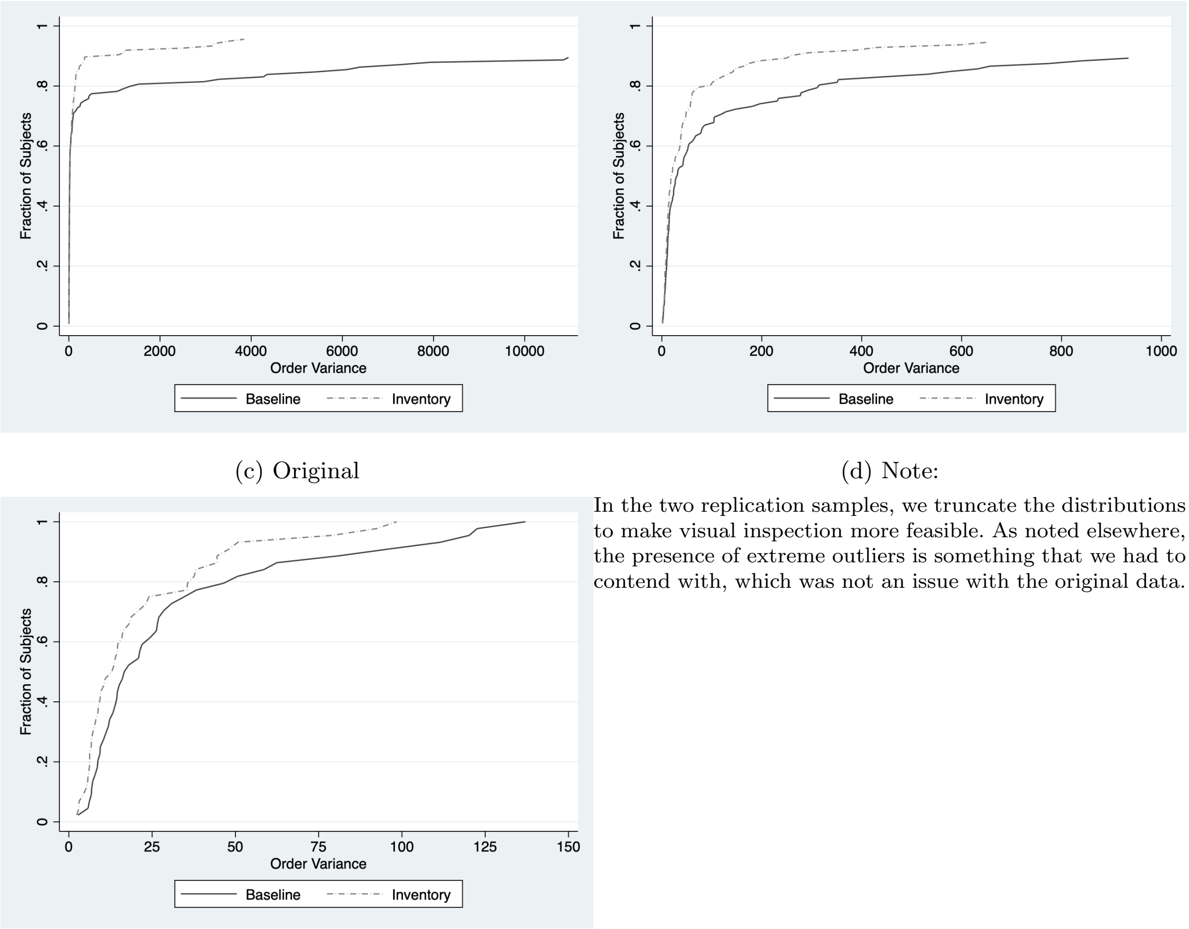
Notes: 1. S.D. stands for standard deviation, while Med. stands for median.
2. We report the Mann-Whitney test under the median column because, under the location-shift assumption, the Mann-Whitney test is a test of equality of medians. We discuss the tenability of this assumption further in the main text. These cells are shaded in grey to indicate that these are the test results being used to judge replication.

References

Croson, Rachel, Karen Donohue. 2006. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science* **52**(3) 323–336.

Hyndman, Kyle, Matthew Embrey. 2018. Econometrics for experiments. Karen Donohue, Elena Katok, Stephen Leider, eds., *Handbook of Behavioral Operations*, chap. 2. Wiley, 35–88.

Figure 1 Empirical CDFs of Data for Replication and Original Samples



(a) UTD

(b) UW-Madison

(c) Original

(d) Note:

In the two replication samples, we truncate the distributions to make visual inspection more feasible. As noted elsewhere, the presence of extreme outliers is something that we had to contend with, which was not an issue with the original data.

Random Demand Realizations

In Table A1, we provide the random demand realizations that were used in the original experiment.

Period	Demand	Period	Demand
1	0	25	5
2	7	26	6
3	4	27	6
4	8	28	8
5	5	29	3
6	3	30	3
7	2	31	8
8	7	32	1
9	8	33	3
10	4	34	2
11	5	35	5
12	1	36	1
13	3	37	3
14	1	38	2
15	1	39	5
16	0	40	4
17	4	41	4
18	1	42	5
19	0	43	6
20	0	44	3
21	2	45	0
22	7	46	8
23	8	47	1
24	2	48	8

Table A1 Random Demand Realizations Used in the Original Experiment