The minimum mean paradox: A mechanical explanation for apparent experiment aversion

Robert Mislavsky[a], Berkeley J. Dietvorst[b], and Uri Simonsohn[c,1]

Meyer et al. (1) propose that people object to "experiments that compare two unobjectionable policies" (their title). In our own work (2), we arrive at the opposite conclusion: People "don't dislike a corporate experiment more than they dislike its worst condition" (our title). In this letter we reanalyze the 7 studies in table 1 of ref. 1, for they most closely resemble ours. We conclude that the evidence for experiment aversion is caused by a statistical artifact.

In those studies, 3 separate groups of people indicated the acceptability of policy A, policy B, or an experiment testing both policies. The average acceptability of the experiment was lower than the average of either policy. This pattern was used as evidence of experiment aversion, but it is actually nondiagnostic.

To illustrate, imagine an experiment giving people a dessert with either dairy (A) or peanuts (B). If 30% of people are lactose-intolerant and another 30% have peanut allergies, A and B are each objectionable to 30% of people. However, the experiment is objectionable to 60%. This higher rate has nothing to do with experimentation. It arises because some object to A, while others object to B. This is why in our studies (2) policies A and B were evaluated by the same people, and we compared the least acceptable policy for each participant to the acceptability of the experiment. In the desserts example, we would compare the share of people objecting to the experiment to the 60% objecting to either A or B, whichever is each individual's worst condition.

In Meyer et al.'s (1) studies, each participant rated only one policy, and thus we cannot determine each participant's worst policy. Instead, we approximated this necessary analysis by pairing observations from the A and B conditions in their data. We treated each resulting pair as if coming from one participant, computing the worst policy for each pair. Because 2 evaluations made by the same person are not independent, we did not form pairs drawing observations independently from A and B. Instead, we first estimated the correlation between evaluations of A and B by asking Amazon Mechanical Turk participants (n = 99) to evaluate all scenarios. We used the resulting average within-scenario, within-person correlation, r = 0.33, as the correlation for the normal variables we drew from. We converted these normal values to quantiles and took such quantiles from Meyer et al.'s A and B samples (see refs. 3 and 4). Fig. 1 shows that, contrary to experiment aversion, experiments were rated almost identically to their worst arms (overall: mean [M] = 3.48 vs. M = 3.44).

Assigning different participants to evaluate different treatments also has psychological consequences (5, 6) that probably increase the appearance of experiment aversion. For example, in studies 1 and 2, in each policy scenario a hospital implements a single, essentially free intervention to reduce infections. However, in the A/B test condition, the hospital identifies 2 interventions and randomly implements only 1. Only in the A/B test condition is a treatment withheld. This confound is avoidable. In the policy scenarios the director could choose 1 of 2 treatments for everyone.

[a]Carey Business School, Johns Hopkins University, Baltimore, MD 21202; [b]Booth School of Business, The University of Chicago, Chicago, IL 60637; and [c]ESADE Business School, Ramon Llull University, 08034 Barcelona, Spain

Author contributions: R.M., B.J.D., and U.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

Data deposition: Data, original materials, and analysis scripts related to this work have been deposited in the Open Science Framework (https://osf.io/jcwnz/).

[1]To whom correspondence may be addressed. Email: uri.sohn@gmail.com.

First published November 6, 2019.

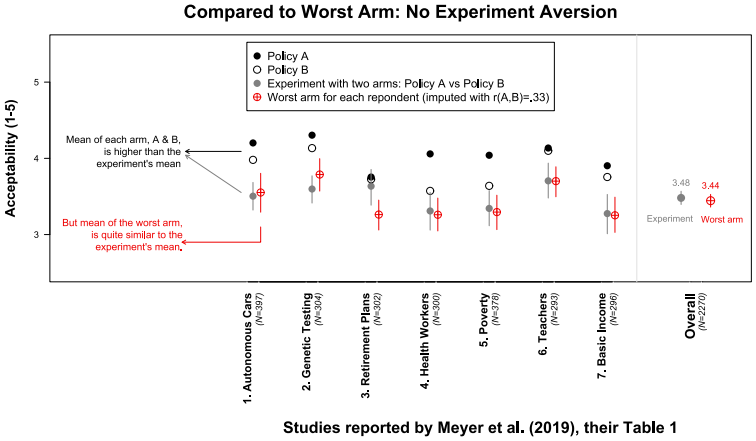

Fig. 1. All analyses were performed on data collected by Meyer et al. (1). Average acceptability of policies A and B (black and white circles, respectively), and of the experiment (gray circle), reproduce results in table 1 of ref. 1. The acceptability of the worst arm is computed by randomly pairing observations between A and B, inducing the correlation within pair that we obtained in new data: r(A,B) = 0.33. For each of the 7 studies, 10,000 pairs (with replacement) were formed. The lowest-rated policy was computed within each pair and averaged across them. This is the estimated average acceptability of the worst arm. Vertical lines depict 95% (bootstrapped) confidence intervals. Note: results are not sensitive to the imputed correlation of r(A,B) = 0.33. With r = 0.5 the mean of the worst arm is M = 3.51, still very similar to the experiment, M = 3.48. With r = 0.2 it is M = 3.40. The R code to reproduce Fig. 1 is available from https://osf.io/jcwnz/ (7).

1 M. N. Meyer *et al.*, Objecting to experiments that compare two unobjectionable policies or treatments. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10723–10728 (2019).
2 R. Mislavsky, B. Dietvorst, U. Simonsohn, Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition. *Mark. Sci.*, 10.1287/mksc.2019.1166.
3 M. C. Cario, B. L. Nelson, Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=B0034903405080939220021 3C01B4D84?doi=10.1.1.48.281&rep=rep1&type=pdf. Accessed 14 October 2019.
4 I. Yahav, G. Shmueli, On generating multivariate Poisson data in management science applications. *Appl. Stochastic Models Data Anal.* **28**, 91–102 (2012).
5 M. H. Birnbaum, How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychol. Methods* **4**, 243–249 (1999).
6 C. K. Hsee, G. F. Loewenstein, S. Blount, M. H. Bazerman, Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychol. Bull.* **125**, 576–590 (1999).
7 R. Mislavsky, B. Dietvorst, U. Simonsohn, Data from "PNAS letter experiment aversion" Open Science Framework. https://osf.io/jcwnz/. Deposited 14 October 2019.