PNAS LETTER

Experiment aversion does generalize, but it can also be mitigated

Randi L. Vogt[a], Patrick R. Heck[a], Duncan J. Watts[b,c,d], Christopher F. Chabris[a,1], and Michelle N. Meyer[a,1,2]

Mazar, Elbaek, and Mitkidis (MEM) assert (1)—in an article edited by Berkeley Dietvorst—that experiment aversion (EA) does not generalize and that Mislavsky, Dietvorst himself, and Simonsohn (2) were essentially correct that EA does not exist. In fact, their data show only that EA can vary by circumstance and be mitigated, as we ourselves suggested (3).

MEM describe experiments making "a small number of changes to the wording of [our] scenarios to further enhance respondents' understanding." Far from "trivial" or "reasonabl[y] minor" (4), these are debiasing interventions (Table 1). Even the new results MEM discuss in their reply to Bas et al. (5)—as well as additional, currently unreported results (6)—suggest that these wording changes yield less negative sentiment toward experiments when debiasing language is present and more negative sentiment when it is absent. Yet as Bas et al. correctly point out, because MEM fail to "systematically and orthogonally manipulat[e]" (5) or test (e.g., by modeling interaction effects) these many changes, nor even to address most of them, it is impossible to know which changes affect participants' judgments about experiments, and to what extent.

For example, we previously identified lack of consent as a partial explanation for EA (3) and experimentally demonstrated that people are less averse to consensual experiments (8). In all MEM studies [except their successful direct replication of Meyer et al. (3)], participants are told "[t]here are 3 hospitals you can choose to be treated at" and the first and primary dependent measure asks, "How likely are you to choose to be treated at this hospital?" That people are less averse to consensual pragmatic trials is unsurprising, not especially actionable [since consent reduces external validity, is typically impractical, and hence is absent from most corporate and many pragmatic healthcare trials (10)], and not evidence that EA fails to generalize.

We similarly previously (3) identified the misbelief that the decision-maker should already know what works best as another explanation for EA. MEM told participants that the interventions "may help" and that "not everybody responds to the treatment with them," indicating that their efficacy is unknown. Unsurprisingly, when their illusion of knowledge is pierced in this way, participants display less EA. In fact, MEM's new studies (4, 6) demonstrate this: When language debiasing the illusion of knowledge was removed, negative sentiments toward experiments increased substantially compared to the effect sizes from the corresponding original studies (in the within-subjects study, 6 percentage points or 28% more participants showed EA; in the between-subjects study, 11 percentage points or 48% more participants rated the experiment as inappropriate).

And as we ourselves noted (3), previous research already showed that describing the same project as a "study" versus an "experiment" can affect perceptions (11). MEM characterize "experiment" as biased language, but such descriptions are the norm: Media report on controversial "experiments" and rarely acknowledge expert uncertainty about the interventions experiments contain.

Even using debiasing materials, MEM's data reveal more EA than the authors acknowledge. In 8 of 9 vignettes, more than 25% of participants ranked the A/B test as the worst option. Moreover, and contrary to MEM's claim—in both their original paper and their reply—that people often prefer experiments, in none of their vignettes (or ours) was there significant experiment appreciation [when correctly defined as the inverse of EA, i.e., preferring the A/B test to the highest (8, 9)—not the lowest (1)—rated policy]. Indeed, in several vignettes, significantly more participants were experiment-averse than experiment-appreciative. When an influential minority fails to appreciate experiments, valuable research may not occur (12, 13).

In our work and others' (13, 14), EA generalizes across domains [medical, public health, public policy, technology; (3, 7–9, 13)], scenarios [safety checklists for catheterization and intubation, prescribing hypertensive and corticosteroid drugs, return of results from genetic testing, retirement savings plans, overrides for autonomous vehicles, ventilator proning for COVID patients, post-COVID school reopening, rules for wearing masks during COVID, distribution of COVID vaccines, recruitment of health workers, poverty alleviation strategies, teacher well-being strategies, basic income plans, lead abatement strategies; (3, 7–9, 13)], populations [laypeople, clinicians, public sector leaders; (3, 9, 13)], and levels of consent [conducting the A/B test after obtaining consent, conducting it without asking for consent, and silence about whether consent was sought; (8)]. MEM themselves show that EA generalizes across seven dependent measures [(in)appropriate, (un)ethical, (ir)responsible, (un)professional, (un)informed, backfire/succeed, likely to choose]. That said, as a social-psychological phenomenon, we should expect EA to vary across settings and societies (e.g., collectivistic versus individualistic) and to be amenable to mitigation. Future research

Author affiliations: [a]Department of Bioethics and Decision Sciences, Geisinger, Danville, PA 17822; [b]Department of Computer and Information Science, University of Pennsylvania, Pennsylvania, PA 19104; [c]Annenberg School of Communication, University of Pennsylvania, Pennsylvania, PA 19104; and [d]Operations, Information, and Decisions Department, University of Pennsylvania, Pennsylvania, PA 19104

Author contributions: R.L.V. analyzed data; and R.L.V., P.R.H., D.J.W., C.F.C., and M.N.M. wrote the paper.

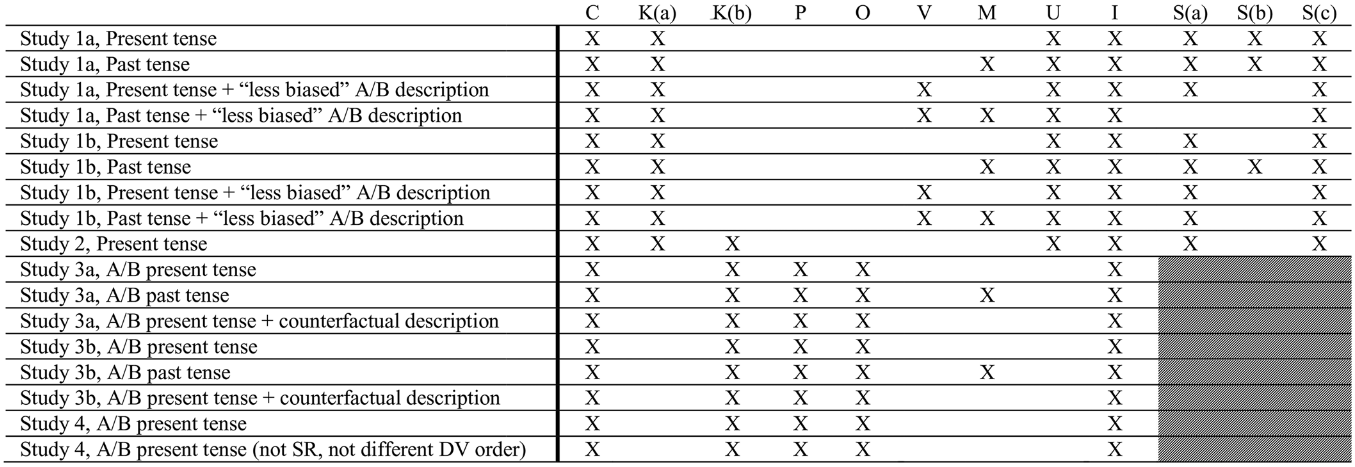
The authors declare no competing interest.

[1] C.F.C. and M.N.M. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: michellenmeyer@gmail.com.

Published May 2, 2024.

Table 1. Select methodological, analytic, and interpretative issues with Mazar et al. (1)



	C	K(a)	K(b)	P	O	V	M	U	I	S(a)	S(b)	S(c)
Study 1a, Present tense	X	X						X	X	X	X	X
Study 1a, Past tense	X	X					X	X	X	X	X	X
Study 1a, Present tense + “less biased” A/B description	X	X				X		X	X	X		X
Study 1a, Past tense + “less biased” A/B description	X	X				X	X	X	X			X
Study 1b, Present tense	X	X						X	X	X		X
Study 1b, Past tense	X	X					X	X	X	X	X	X
Study 1b, Present tense + “less biased” A/B description	X	X				X		X	X	X		X
Study 1b, Past tense + “less biased” A/B description	X	X				X	X	X	X	X		X
Study 2, Present tense	X	X	X					X	X	X		X
Study 3a, A/B present tense	X		X	X	X				X	▨	▨	▨
Study 3a, A/B past tense	X		X	X	X		X		X	▨	▨	▨
Study 3a, A/B present tense + counterfactual description	X		X	X	X				X	▨	▨	▨
Study 3b, A/B present tense	X		X	X	X				X	▨	▨	▨
Study 3b, A/B past tense	X		X	X	X		X		X	▨	▨	▨
Study 3b, A/B present tense + counterfactual description	X		X	X	X				X	▨	▨	▨
Study 4, A/B present tense	X		X	X	X				X	▨	▨	▨
Study 4, A/B present tense (not SR, not different DV order)	X		X	X	X				X	▨	▨	▨

Notes: X indicates that Mazar et al. (1) has the issue listed in the column header associated with the key below. Shaded cells indicate that this critique is not applicable due to the between-subjects design of the study.

Key:

C: *Experiment framed as consensual.*

Asking participants to “choose” whether to visit a hospital that uses policy A, policy B, or conducts an A/B test does not measure attitudes toward pragmatic (nonconsensual) RCTs—and contaminates all remaining DVs, e.g., (in)appropriate, (ir)responsible, (un)informed.

K: *Illusion of knowledge weakened.*

(a) Telling participants that the decision-maker thinks the policy interventions “may help” debiases their tendency to think that the decision-maker already knows which policy is best (and should implement that without an A/B test).

(b) Telling participants that “not everybody responds to the treatment with them” increases uncertainty about the wisdom of, and what is known about, policies A and B.

P: *Policy arms made less palatable.*

Describing the decision maker as “randomly decid[ing]” which policy to implement (only in the A and B conditions) suggests policies are chosen without care and thought.

O: *Oversight added.*

By changing the decision maker from one individual doctor to a hospital/clinic director, oversight is implicitly added which makes the experiment seem more legitimate, and thus more palatable to participants.

V: *Experiment description lacks external validity.*

Using the “less biased” language of “test” in lieu of “experiment” is not representative of the language typically used to discuss A/B tests in the media.

M: *Measuring preference for evidence-based medicine.*

When—in a “past tense” vignette—participants “choose” to be treated at a hospital where the “director assessed which drug, A or B, had had the best outcomes for their patients, and from then on, all new patients...are prescribed that drug,” this shows that people are willing to free ride on past A/B tests to receive evidence-based treatments, not that EA fails to generalize.

U: *Underpowered.*

Sample size (N ≈ 135 to 155 per variation) is not large enough to detect experiment aversion [power analysis by Heck et al. (7, p. 18949) recommended N ≈ 300 to 450]. No power analyses were reported by Mazar et al. (1).

I: *Inadequate evidence for claims.*

Claims about differences in experiment aversion caused by changes in tense, language, the emphasis of the counterfactual, and the order of questions require testing for interaction effects and recruiting substantially larger samples, neither of which were done.

S: *Unjustified conclusion that “people either significantly prefer experiments or do not significantly differentiate between them and the universal implementation of the individual policies.”*

(a) >25% of participants ranked the A/B test worst.

(b) A significantly greater proportion of participants are experiment-averse than are experiment-appreciative.

(c) When defined as the difference between the rating of the A/B test and the rating of the highest-rated policy (the inverse of experiment aversion), there is no significant experiment appreciation. Mazar et al. (1) report finding “experiment preference” in several studies. They define experiment preference as the opposite of experiment aversion—that is, the difference between the rating of the A/B test and the rating of the lowest-rated policy. Using this definition, a person shows experiment “preference” if they rate the A/B test higher than they rate their least-preferred policy. We believe that such a “preference” is not meaningful and instead calculate experiment appreciation which is defined as the difference between the rating of the A/B test and the rating of the highest-rated policy (8, 9). Using this definition, a person shows experiment appreciation when they like the A/B test more than their favorite policy, a characteristic that we find meaningful.

should abandon attempts to “disprove” experiment aversion and instead focus on when and why it happens, and how to make it happen less.

ACKNOWLEDGMENTS. Supported by Office of the Director, NIH (3P30AG034532-13S1 to M.N.M. and C.F.C.) and funded by the Food and Drug Administration (FDA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the FDA. Also supported by Riksbankens Jubileumsfond grant “Knowledge Resistance: Causes, Consequences, and Cures” to Stockholm University (to C.F.C.), via a subcontract to Geisinger Health System. We thank Rebecca Mestechkin for excellent research assistance. Analysis code has been deposited in OSF (https://osf.io/jn789/).

1. N. Mazar, C. T. Elbaek, P. Mitkidis, Experiment aversion does not appear to generalize. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2217551120 (2023).
2. R. Mislavsky, B. Dietvorst, U. Simonsohn, Critical condition: People don’t dislike a corporate experiment more than they dislike its worst condition. *Marketing Sci.* **39**, 1092–1104 (2020).
3. 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).
4. N. Mazar, C. T. Elbaek, P. Mitkidis, Reply to Bas et al.: The difference between a genuine tendency and a context-specific response. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2318010120 (2023).
5. B. Bas, J. Vosgerau, R. Ciulli, No evidence that experiment aversion is not a robust empirical phenomenon. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2317514120 (2023).
6. N. Mazar *et al.*, Data from “Replies to Letters about Experiment Aversion”. OSF. https://osf.io/whz3b/. Accessed 21 December 2023.
7. P. R. Heck, C. F. Chabris, D. J. Watts, M. N. Meyer, Objecting to experiments even while approving of the policies or treatments they compare. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 18948–18950 (2020).
8. R. L. Vogt, R. M. Mestechkin, C. F. Chabris, M. N. Meyer, Objecting to consensual experiments even while approving of nonconsensual imposition of the policies they contain. PsyArXiv [Preprint] (2023a). https://doi.org/10.31234/osf.io/8r9p7 (Accessed 15 September 2023).
9. R. L. Vogt *et al.*, Experiment aversion among clinicians and the public–an obstacle to evidence-based medicine and public health. MedRxiv [Preprint] (2023b). https://www.medrxiv.org/content/10.1101/2023.04.05.23288189v1 (Accessed 15 September 2023).
10. L. I. Horwitz, M. Kuznetsova, S. A. Jones, Creating a learning health system through rapid-cycle, randomized testing. *N. Engl. J. Med.* **381**, 1175–1179 (2019).
11. S. J. Cico, E. Vogeley, W. J. Doyle, Informed consent language and parents’ willingness to enroll their children in research. *IRB* **33**, 6–13 (2011).
12. V. Prasad, “3.17 COVID-19 and schools in Norway with Dr. Atle Fretheim & cancer biology with Dr. Anthony Letai,” *Plenary Session [podcast]*, https://www.plenarysessionpodcast.com/episodes/x3846t9sxwywced-slyf3-rstp6-685wr-sc7gw-fklz9. Accessed 15 September 2023.
13. E. Cardon, L. Lopoo, Randomized controlled trial aversion among public sector leadership: A survey experiment. *Eval. Rev.* 0193841X231193483 (2023), https://doi.org/10.1177/0193841X231193483.
14. B. Bas, R. Ciulli, J. Vosgerau, Why do people condemn and appreciate experiments? in *Proceedings of the 51st Annual European Marketing Academy Conference* (Budapest, 2022).