PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Experiment aversion does not appear to generalize

Nina Mazar[a,1], Christian T. Elbaek[b], and Panagiotis Mitkidis[b]

Edited by Berkeley J. Dietvorst, University of Chicago, Chicago, IL; received October 24, 2022; accepted March 8, 2023 by Editorial Board Member Dalton Conley

Over the past decade, governments and organizations around the world have established behavioral insights teams advocating for randomized experiments. However, recent findings by M. N. Meyer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 116, 10723–10728 (2019) and P. R. Heck, C. F. Chabris, D. J. Watts, M. N. Meyer, *Proc. Natl. Acad. Sci. U.S.A.* 117, 18948–18950 (2020) suggest that people often rate randomized experiments as less appropriate than the policies they contain even when approving the implementation of either policy untested and when none of the individual policies is clearly superior. The authors warn that this could cause policymakers to avoid running large-scale field experiments or being transparent about running them and might contribute to an adverse heterogeneity bias in terms of who is participating in experiments. In one direct and six conceptual preregistered replications (total N = 5,200) of the previously published larger-effect studies, using the same main dependent variable but with variations in scenario wordings, recruitment platforms, and countries, and the addition of further measures to assess people's views, we test the generalizability and robustness of these findings. Together, we find that the original results do not appear to generalize. That is, our triangulation reveals insufficient evidence to conclude that people exhibit a common pattern of behavior that would be consistent with relative experiment aversion, thereby supporting recent findings by R. Mislavsky, B. Dietvorst, U. Simonsohn, *Mark. Sci.* 39, 1092–1104 (2020). Thus, policymakers may not need to be concerned about employing evidence-based practices more so than about universally implementing policies.

replication | policy | nudging | behavioral science practice | randomized controlled trial

Significance

Governments and organizations worldwide have established behavioral insights teams advocating for randomized experiments as a critical tool to advancing human welfare. Recently, however, researchers published empirical evidence to conclude that people tend to find randomized experiments as less appropriate than the universal implementation of its underlying unobjectionable policies and warn that this common pattern may threaten evidence-based practices and policy. Our conceptual and direct replications of the previously published larger-effect studies in support of experiment aversion indicate that the evidence for this conclusion is neither generalizable nor robust. That is, our findings suggest that there is not enough evidence to conclude that there exists a general pattern such as relative experiment aversion.

Over the past decade and more, local and central governments, international organizations, as well as for profit and not-for-profit organizations around the world have witnessed the establishment of behavioral insights teams (e.g., refs. 1–3). These units have been advocating for evidenced-based policy—relying on the gold-standard: randomized controlled trials (RCTs; (4)) – as critical for advancing human welfare. However, recent works by Meyer et al. (5, 6) and Heck et al. (7) raise concerns about a significant barrier to evidence-based practice: relative *Experiment Aversion*. Specifically, the authors assessed attitudes toward A/B test implementations in contexts where the individual policies are unobjectionable, and neither is found to be clearly superior and conclude that people are often paradoxically more approving of policies or treatments (A or B) being universally implemented than of randomized experiments (A/B tests) to determine which of those individual policies or treatments is superior.

The conclusion put forward by the recent work (5–7) has potentially wide-reaching negative implications for evidence-based policy. As the authors argue, individuals' attitudes toward experiments matter not because these individuals decide on policy implementation, but because they affect policymakers' decisions; policymakers who anticipate objections toward experiments "may forgo them in favor of universal implementation or may conduct randomized evaluations in secret, neither of which is optimal" (ref. 5, p. 10727).

The evidence for relative experiment aversion published by Meyer et al. (5, 6) and Heck et al. (7) appears "robust and general" and often "substantial" (ref. 5, p. 10725). Specifically, Meyer et al. (5) present persistent results across 16 between-subject design studies (i.e., each participant learns about and evaluates only one approach: either the experiment or the universal implementation of one of the two individual policies) with 5,873 participants spanning nine scenario domains. Fourteen of these studies revealed significant ($P < 0.05$) experiment aversion—defined as people rating an agent's decision for an experiment as less appropriate than the pooled individual policy A and individual policy B conditions mean (A/B Effect: Mean (A/B test) < Mean (A+B pooled))—and two studies revealed no significant difference.*

*What previous research (5–7) defines as *experiment aversion* does not indicate that A/B test decisions are generally rated as inappropriate in absolute terms (i.e., mean ratings below the midpoint of the bipolar scale). In fact, the average appropriateness ratings of A/B tests in each of our studies were above the scales' midpoints, which is consistent with previous results (5–9).

Author contributions: N.M. and P.M. designed research; N.M. and P.M. performed research; C.T.E. analyzed data; and N.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. B.J.D. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: nina@ninamazar.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2217551120/-/DCSupplemental.

Published April 10, 2023.

In addition, a subset of the original authors, Heck et al. (7), present persistent results across five within-subject design studies (i.e., each participant learns about and evaluates all the three types of approaches: the experiment and the universal implementations of each of the two individual policies), with 1,955 participants spanning five of the original nine scenarios (5). All five studies reveal significant ($P < 0.05$) experiment aversion as defined in ref. 5. The later joint evaluation design also allows to test and refute the concern that people may object to one or both of the individual policies and therefore may be rationally rating the experiment as they would rate the universal implementation of their least-preferred of the two individual policies, as suggested by Mislavsky et al. (8). Mislavsky et al.'s (8) suggestion was based on their examination (9) of the acceptability of experiments in relatively low-stake corporate marketing settings such as discount implementations in a ride-sharing scenario and product recommendation system changes in an Amazon.com scenario. There, across six studies, Mislavsky et al. (9) provide evidence in support of a "critical condition" account of experiment evaluation: Experiments are rated as at least as acceptable as the least acceptable policies they contain. However, contrary to Mislavsky et al.'s (8, 9) account, when applying the more stringent testing criterion (i.e., A/B Effect: Mean (A/B test) < Min (A, B)) to their joint-evaluation studies, Heck et al. (7) continue to find significant ($P < 0.05$) evidence for experiment aversion in four of the five studies. That is, their findings are in opposition to Mislavsky et al.'s (9) critical condition account and instead support Meyer et al.'s (ref. 6, p. 23885) conclusion that relative experiment aversion "cannot be explained away as an 'artifact' or 'confound'" of participants mechanically rating the experiment as they would its worst treatment arm.

Finally, Meyer et al. (5) and Heck et al. (7) find experiment aversion to persist across studies not only when counterfactual options are highlighted (as e.g., in the joint-evaluation design) but even when scenario patients are described as being treated unequally and/or randomly also in the individual policy conditions, and not only in the A/B test condition. Together, the results lead the authors to maintain their conclusion that there exists a "genuine aversion to randomized evaluation" among people (ref. 7, p. 18948).

Given the potential implications of the existence of an "anomalous" (ref. 7, p. 18948) experiment aversion as put forward by Meyer et al. (5, 6) and Heck et al. (7), the goal of our research was to use triangulation (10, 11) to examine whether we could replicate their evidence after changes that might be viewed as trivial, considering the persistence of the previously published results (5–7). In particular, we made a small number of changes to the wording of the original scenarios to further enhance respondents' understanding of the presented options (e.g., making it clearer that the agent does not know which individual policy is best/better) and to describe the scenarios in self-relevant terms to make the hypothetical more tangible and produce potentially more truthful responses (*SI Appendix*, Tables S1–S3 present the side-by-side comparison of the language of the original scenarios from refs. 5 and 7 and our adapted ones). Furthermore, we added a variety of questions to more thoroughly assess people's attitudes toward experiments such as how likely participants would be to choose to be exposed to each of the policies (typically our first question), and, after asking about the appropriateness of the agent's decision (i.e., the original DV), we inquired about five more specific dimensions with respect to the agent's decision: the extent to which it was viewed as ethical, responsible, professional, informed, and likely to succeed (each on bipolar Likert-type response scales).

In addition to these changes, we tested within studies the sensitivity of results to variations in i) tense, to examine whether respondents' views about experiments may differ based on if an experiment was imminent (present tense as done exclusively in refs. 5 and 7) versus completed and the policy with the best outcome implemented (i.e., past tense), ii) variations in language to describe the experiment with simpler and less presumably biased words (e.g., replacing the use of "experiment" with "test"), iii) variations in the extent to which the experiment description emphasized the counterfactual option of randomly implementing one of the two individual policies universally without testing, and iv) order of questions and whether the scenarios were framed in self-relevant terms.

Finally, we varied between studies the platforms and Western countries from which we recruited participants to examine the generalizability of findings across populations. The originally published studies (5, 7) did not vary any factors between participants (other than A, B, A/B test), and were exclusively conducted in the United States, primarily with participants from Amazon Mechanical Turk (MTurk).

We conservatively based our studies on what Meyer et al. (ref. 6, p. 23885) describe to be their three "larger-effect" scenarios: *Safety-Checklist*, *Best Drug* (adapted to the topical COVID-19 pandemic before vaccines were available[†]), and *Best Drug: Walk-In Clinic* (described as identical to the Best Drug scenario, except that the prescription of Drug A vs. Drug B is "essentially random"; ref. 5, *SI Appendix*, p. 6). In fact, Heck et al. (7) describe the Safety-Checklist and the Best Drug: Walk-In Clinic scenarios to be "two of the most-studied and important domains" (p. 18948).

Together, for our triangulation, we ran seven preregistered studies ($N = 5{,}200$). Specifically, we conducted six conceptual replications and one direct replication with 2.5× the original sample size (12), on three different platforms (CloudResearch to access MTurk, Lucid Theorem, and Prolific Academic) with a multitude of Western national cultures (the United States, the United Kingdom, Scandinavian countries, Germany, and Austria). For an overview of our studies (three joint evaluation studies as done in ref. 7 and four separate evaluation studies as done in ref. 5), and their major differences from the original studies, see Tables 1 and 2.

Results

Tables 1 and 2, second last column, present the A/B Effect for the original DV from Meyer et al. (5) and Heck et al. (7): Appropriateness of the agent's decision (from very inappropriate to very appropriate), for each of our conditions. The "–" and "+" symbols represent a significant relative aversion and preference, respectively, for experiments, and "0" represents no significant difference, all based on $P < 0.05$ and adjusted for multiple comparisons. As can be seen, in our six conceptual replications, we find that people either significantly prefer experiments or do not significantly differentiate between them and the universal implementation of the individual policies, deviating dramatically from the previously published empirical evidence (5, 7). That is, only in our direct replication Study 5 do we replicate a significantly negative and large effect size ($d = -0.84$, $P < 0.001$), indicating what Meyer et al. (5) define as experiment aversion. However, when we vary the approach as done in our conceptual replications, we often find the opposite pattern with significantly positive effects, indicating a relative *preference for experiments* ranging in size from small ($d = 0.21$, $P < 0.05$) to medium ($d = 0.44$, $P = 0.001$).

In addition, previous research has documented that people's attitudes can be shifted around based on, for example, the type of

[†]Because a COVID-19 scenario presented a sufficiently real possibility to participants at that time, we thought our adaption would increase the external validity and generalizability of the scenario study.

Table 1. Overview of our studies with joint evaluations of A, B, A/B test

Study	Scenario	Sample (n, platform, country, # between-ss conditions, date conducted)	Major deviations from most relevant originally published studies	Between-ss conditions	Original DV: Agent's decision is (in)/appropriate: A/B test M* (SD)	Original DV: Agent's decision is (in)/appropriate: A/B effect†	Selected other DVs: A/B effect† for likelihood to choose; decision is (un)/ethical; decision is likely to (backfire)/be successful.
1a	Safety Checklist	N = 550, Scandinavia, Prolific, 4 conds (2 × 2), 1/2020	From ref. 7, Study 1: • Country • Options not counterbalanced • SR (self-relevant) • DV order • Platform	Present tense	4.36 (1.78)	**0** t = 0.08, d = 0.01, *P* = 1	**0 − +**
				Past tense	4.53 (1.85)	**0** t = 1.02, d = 0.12, *P* = 1	**+ − +**
				Present tense + less biased A/B description	4.84 (1.73)	**0** t = 1.67, d = 0.20, *P* = 0.392	**+ 0 +**
				Past tense + less biased A/B description	5.07 (1.77)	**+** t = 3.67, d = 0.44, *P* = 0.001	**+ 0 +**
1b	Safety Checklist	N = 606, U.S., Lucid, 4 conds (2 × 2), 4/2020	From ref. 7, Study 1: • SR • DV order • Platform	Present tense	4.77 (1.93)	**+** t = 2.94, d = 0.21, *P* = 0.015	**+ 0 +**
				Past tense	4.68 (1.88)	**+** t = 2.81, d = 0.21, *P* = 0.023	**+ 0 +**
				Present tense + less biased A/B description	4.62 (1.90)	**+** t = 4.22, d = 0.26, *P* < 0.001	**0 + +**
				Past tense + less biased A/B description	4.93 (1.96)	**+** t = 4.26, d = 0.28, *P* < 0.001	**+ + +**
2	Best Drug	N = 155, UK, Prolific, 1 cond, 4/2020	From ref. 5, Study 4: • Country • COVID context • Joint evaluation with counterbalanced options • SR • DV order • Platform	Present Tense	5.23 (1.56)	**0** t = −0.98, d = −0.06, *P* = 0.33	**0 − +**

Notes: **−** = Relative experiment aversion, **0** = Experiment neutral, **+** = Relative experiment preference; based on $P < 0.05$, adjusted for multiple comparisons (Bonferroni). Gray-shaded row highlights conceptual replication most similar to originally published studies.
*The originally published studies use a 5-point bipolar Likert-type scale (from 1: very inappropriate to 5: very appropriate), we use a 7-point bipolar Likert-type scale.
†Same as Heck et al. (7), we apply the more stringent criterion as suggested by Mislavsky et al. (8): Mean (A/B test) < Min (A, B).

questions asked and their order and type of response scales, and that people's preferences are context dependent (see e.g., refs. 13 and 14 on measurement of attitudes; (15) on joint vs. separate evaluations and thus, the amount of information about forgone alternatives; (16) on constructed preferences). However, none of our additional between-subject factors (e.g., tense: present vs. past) robustly moderated the A/B Effect (see ref. 17).

Finally, Tables 1 and 2, last column, present A/B Effect results for three of our additional measures: one self-relevant measure (likelihood to choose to be treated at the hospital on a scale from not likely at all to very likely) and two of the five measures reflecting the more specific dimensions along which participants were asked to rate the agents' decisions. The latter two were chosen to demonstrate the variance of A/B Effects as they represent the two dimensions that most often resulted in significantly negative or positive A/B Effects. Respondents were most often significantly negative of an agent's decision to experiment relative to universally implementing the individual policies when judging the decision's ethicality, and most often significantly positive (and never significantly negative) when judging its likelihood to be successful and result in reduced infections. But even along the most critical dimension, ethicality of the agent's decision, we found no general tendency toward relative experiment aversion. Instead, participants sometimes judged the agent's decision to experiment as relatively more ethical, sometimes as relatively less ethical, and sometimes they were indifferent between the ethicality of the decision to experiment and the ethicality of the decision to universally implement the individual policies without testing (based on adjusted $P < 0.05$). For an overview of the results across all our measures in each of our studies, see *SI Appendix*, Fig. S1 and Tables S4–S10. For extended study notes, see *SI Appendix*, Supporting Information Text.

Discussion

Our research reveals that even consistently replicating results as in refs. 5–7 and our Study 5 does not necessarily mean that those results represent robust insights. Instead, they may merely be artifacts or confounds of contextual circumstances and insufficient triangulation with multiple approaches (7, 10). Focusing on the original dependent variable: Appropriateness of the agent's decision (Tables 1 and 2, second last column) and comparing the previous work to our two conceptual replication studies most similar to the

Table 2. Overview of our studies. Separate evaluations of A present tense, B present tense, and variations of A/B test

Study	Scenario	Sample (n, platform, country, # between-ss conditions, date conducted)	Major deviations from most relevant originally published study: ref. 5, Study 5a	Between-ss conditions	Original DV: Agent's decision is (in)/ appropriate — A/B test M* (SD)	Original DV: Agent's decision is (in)/ appropriate — A/B effect†	Selected other DVs — A/B effect‡ for likelihood to choose; decision is (un)/ethical; decision is likely to (backfire)/be successful.
3a	Best Drug: Walk-In Clinic	N = 1,418, United States, CloudResearch (MTurk), 5 conds, 8/2020	• SR • DV order • Decision agent = Hospital director – akin to Best Drug scenario (i.e., ref. 5, Study 4)	A/B present tense	4.04 (1.95)	**0** t = 0.22, d = 0.02, *P* =1	**− 0 0**
				A/B past tense	4.20 (1.87)	**0** t = 1.48, d = 0.11, *P* = 0.85	**0 0 +**
				A/B present tense + counter-factual description	4.19 (1.88)	**0** t = 1.35, d = 0.10, *P* = 1	**0 0 0**
3b	Best Drug: Walk-In Clinic	N = 781, Germany & Austria, Prolific, 5 conds, 11/2020	• Country • SR • DV order • Decision agent = Hospital director – akin to Best Drug scenario (i.e., ref. 5, Study 4) • Platform	A/B present tense	4.56 (1.66)	**+** t = 3.83, d = 0.39, *P* < 0.001	**0 0 +**
				A/B past tense	4.44 (1.57)	**+** t = 3.25, d = 0.33, *P* = 0.01	**0 0 +**
				A/B present tense + counter-factual description	4.38 (1.57)	**+** t = 2.85, d = 0.29, *P* = 0.030	**0 0 0**
4‡	Best Drug: Walk-In Clinic	N = 898, United States, Prolific, 6 conds (2 × 3), 12/2020	• SR • DV order • Decision agent = Hospital director – akin to Best Drug scenario (i.e., ref. 5, Study 4) • Platform	A/B present tense	4.68 (1.78)	**+** t = 3.69, d = 0.37, *P* < 0.001	**0 + +**
				A/B present tense (not SR, not different DV order)	4.84 (1.91)	**+** t = 3.91 d = 0.40, *P* < 0.001	**0 + +**
5	Best Drug: Walk-In Clinic	N = 792, United States, CloudResearch (MTurk), 3 conds, 4/2021	DIRECT REPLICATION	A/B present tense (not SR, not different DV order)	3.54 (1.34)	**−** t = −9.75, d = −0.84, *P* < 0.001	**− − −**

Notes: **−** = Relative experiment aversion, **0** = Experiment neutral, **+** = Relative experiment preference; based on $P < 0.05$, adjusted for multiple comparisons (Bonferroni). Gray-shaded row highlights conceptual replication most similar to originally published study.
*The originally published studies use a 5-point bipolar Likert-type scale (from 1: very inappropriate to 5: very appropriate), we use a 7-point bipolar Likerrt-type scale in all but our Study 5.
†Same as Meyer et al. (5), we apply as criterion Mean (A/B test) < Mean (A+B pooled).
‡This study has a fully crossed design and therefore compares A/B present tense to its corresponding (i.e., with or without SR + different DV order) A present tense and B present tense conditions.

previously published research (our joint evaluation Study 1b and our separate evaluation Study 4; in particular, see gray-highlighted condition rows) suggest that the choice of recruitment platform may have contributed to the differing results. In fact, the sensitivity to recruitment platforms is already somewhat apparent in ref. 5, where the large-sized experiment aversion observed with respondents from Amazon's Mechanical Turk in its Study 5a (d = –0.64, $P < 0.001$) is reduced to a nonsignificant small effect with US mobile device users recruited from Pollfish in its Study 5b (d = –0.15, $P = 0.056$). And indeed, previous work has documented platform differences among others in participants' attention, comprehension, reliability, and behavior (e.g., refs. 18, 19). It is also possible that what one would assume to be minor tweaks in scenario language given the persistence of the previously published results (such as a small number of changes to further enhance respondents' understanding of the presented options) contributed to our differing findings, which would further undermine the proposition that there exists a general and robust pattern of behavior representing relative experiment aversion.

Our research additionally reveals that as part of triangulation, it may be useful to consider various operationalizations of assessing people's views, the latter of which is particularly important when there is no obvious one correct measure, and the goal is to establish whether there exists a genuine tendency in people's responses. For example, other justifiable ways to measure views toward experiments include asking participants to consider how likely they themselves would be to choose to be exposed to each of the policies (as we have done) and eliciting participants' willingness to publicly vote for an approach or support it financially. In addition, Meyer et al. (5) interpret their findings, which are based on respondents' ratings of the appropriateness of an agent's decision, as expressions of moral concerns (see p. 10727). However, we find that ratings of more nuanced dimensions such as the ethicality of the agent's decision are viewed as somewhat different from the appropriateness question.

In sum, our findings indicate that there appears to be no generalizable or robust empirical evidence for the existence of relative experiment aversion (i.e., a general pattern whereby people anomalously tend to judge a randomized experiment comparing two unobjectionable policies, neither of which is known to be clearly superior, as worse than simply implementing its individual policies[‡]). Our findings are in opposition to those of Meyer et al. (5, 6) and Heck et al. (7) and instead closely align with those of Mislavsky et al.'s (9) critical condition account. Thus, we arrive at the same conclusion as Mislavsky et al. (9) did for companies: Policymakers may not need to be concerned about using evidence-based practice more so than about universally implementing individual policies. In addition, we conclude that whenever communicating a policy, it is useful to remember that there is no "neutral" language, and policymakers can and should be mindful of their communication choices and perhaps even test them.

Materials and Methods

Preregistration. All studies were preregistered. Any deviations from the preregistrations are explained in *SI Appendix*, Supporting Information Text, points #7 and #8.

For consistency and ease of comparability to the original work (5–7), we opted to present the results in terms of the "A/B Effect" as done in refs. 5 and 7. That is, for our joint evaluation studies, we present the more stringent criterion: the comparison between the A/B test mean and the individual worst arm [as proposed by Mislavsky et al. (8) and done in Heck et al. (7)]. For our separate evaluation studies, we present the original criterion: the comparison between the A/B test mean and the pooled individual policy A and individual policy B condition mean [as done in Meyer et al. (5)]. However, we did not preregister these particular A/B Effect comparisons. We preregistered to run critical pairwise comparisons between the conditions. The results of the latter preregistered critical pairwise comparisons, that is, of the A/B test vs. the individual policy A and of the A/B test vs. the individual policy B, are reported in detail in ref. 17 and the takeaways are essentially the same.

Study Format. The scenario texts used in our studies and how they differed from the original ones are presented in *SI Appendix*, Tables S1–S3. Our seven studies were conducted from January 2020 to April 2021 and are presented in the manuscript in the order in which they were conducted over time. Studies 1a and 1b employ a mixed-factorial design. Study 2 employs a within-subject design. Studies 3, 4a, 4b, and 5 employ a between-subject design. Study 5 was a direct replication of Study 5a in Meyer et al. (5). Same as in the original studies, all our studies were Qualtrics surveys and included a series of demographic and other questions subsequent to the evaluation of participants' views toward the policy implementations, such as a science literacy scale, which are not analyzed further in this manuscript.

All protocols were either determined to be exempt from review by the Aarhus University's Research Ethics Committee or covered by Boston University's Institutional Review Board.

Participants. Participants (N = 5,200) signed informed consent forms, were randomly assigned to conditions, and were paid standard participation fees.

Results

Additional notes and results are provided in the *SI Appendix*. The complete analysis is presented in ref. 17. The test statistics of the pairwise comparisons are adjusted for multiple comparisons (Bonferroni), and we only focus on outcomes with adjusted p-values significant at the 5% level. The programming scripts for data management and statistical analyses were written in the statistical environment *R* (version 4.0.3). No conditions or variables were dropped from any analyses we report. All participants, who completed the surveys, were included in the main analysis.

Data, Materials, and Software Availability. All materials including survey instruments, recruitment information, preregistrations, the analysis code, anonymized data, and supplementary results have been shared and made publicly available online as an OSF project at https://osf.io/wqrkv/ (17).

ACKNOWLEDGMENTS. We thank the editors and the three anonymous reviewers for their constructive feedback and support.

‡Looking at "% rating inappropriate" [as done in the figures of Meyer et al. (5) and Heck et al. (7); i.e., sum of participants indicating ratings below the midpoint], we arrive at the same conclusion, see corresponding graphs on cumulative frequencies in ref. 17.

Author affiliations: [a]Department of Marketing, Questrom School of Business, Boston University, Boston, MA 02115; and [b]Department of Management, School of Business and Social Sciences, Aarhus University, Aarhus V 8210, Denmark

1. Z. Afif, W. W. Islan, O. Calvo-Gonzalez, A. Dalton, *Behavioral Science Around the World: Profiles of 10 Countries (English). eMBeD Brief* (World Bank Group, Washington, D.C., 2018).
2. L. A. Manning, A. Dalton, Z. Afif, R. Vakis, F. Naru, *Behavioral Science Around the World: Volume Two–Profiles of 17 International Organizations (English). eMBeD Brief* (World Bank Group, Washington, D.C., 2020).
3. OECD, *Behavioural Insights and Public Policy: Lessons from Around the World* (OECD Publishing, Paris, 2017).
4. R. H. Thaler, C. R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Yale University Press, 2008).
5. 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).
6. M. N. Meyer *et al.*, Reply to Mislavsky et al.: Sometimes people really are averse to experiments. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23885–23886 (2019).
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. Mislavsky, B. J. Dietvorst, U. Simonsohn, The minimum mean paradox: A mechanical explanation for apparent experiment aversion. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 23883–23884 (2019).
9. R. Mislavsky, B. Dietvorst, U. Simonsohn, Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition. *Mark. Sci.* **39**, 1092–1104 (2020).
10. M. R. Munafò, G. D. Smith, Repeating experiments is not enough. *Nature* **553**, 399–401 (2018).
11. T. Yarkoni, The generalizability crisis. *Behav. Brain Sci.* **45**, E1 (2022).
12. U. Simonsohn, Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015).
13. L. R. Fabrigar, J. A. Krosnick, B. L. MacDougall, "Attitude measurement: Techniques for measuring the unobservable" in *Persuasion: Psychological Insights and Perspectives*, T. C. Brock, M. C. Green, Eds. (Sage, 2005), pp. 17–40.
14. J. A. Krosnick, C. M. Judd, B. Wittenbrink, "The measurement of attitudes" in *The Handbook of Attitudes*, D. Albarracín, B. T. Johnson, M. P. Zanna, Eds. (Lawrence Erlbaum Associates, 2005), pp. 21–76.
15. 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).
16. C. R. McKenzie, S. Sher, L. M. Leong, J. Müller-Trede, Constructed preferences, rationality, and choice architecture. *Rev. Behav. Econ.* **5**, 337–360 (2018).
17. N. Mazar, C. T. Elbaek, P. Mitkidis, Online archive for "Experiment aversion does not appear to generalize." Open Science Framework. https://osf.io/wqrkv/. Deposited 30 March 2022.
18. E. Peer, D. Rothschild, A. Gordon, Z. Evernden, E. Damer, Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **54**, 1643–1662 (2022).
19. K. A. Thomas, S. Clifford, Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.* **77**, 184–197 (2017).