Objecting to experiments even while approving of the policies or treatments they compare

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

[a]Center for Translational Bioethics and Health Care Policy, Geisinger Health System, Danville, PA 17822; [b]Autism and Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA 17837; and [c]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA 19104

Edited by Margaret Levi, Stanford University, Stanford, CA, and approved July 4, 2020 (received for review May 13, 2020)

We resolve a controversy over two competing hypotheses about why people object to randomized experiments: 1) People unsurprisingly object to experiments only when they object to a policy or treatment the experiment contains, or 2) people can paradoxically object to experiments even when they approve of implementing either condition for everyone. Using multiple measures of preference and test criteria in five preregistered within-subjects studies with 1,955 participants, we find that people often disapprove of experiments involving randomization despite approving of the policies or treatments to be tested.

field experiments | A/B tests | randomized controlled trials | research ethics | pragmatic trials

Randomized, controlled trials (RCTs)—sometimes known as A/B tests, field experiments, or pragmatic trials—are considered the "gold standard" for evidence in medicine. They are also increasingly relied on in the social sciences (1)—the 2019 Nobel Prize in Economics was awarded to three researchers for their poverty reduction RCTs. Yet people sometimes object to RCTs across domains including medicine, law, economic development, digital platforms, and even public health emergencies like the coronavirus disease 2019 pandemic (2–7).

Recently, Meyer et al. (3) found that people rated A/B tests as less appropriate than an average rating of universally implementing A or B (the "A/B Effect"). There, participants evaluated only one of the three possibilities (policy A, policy B, or the A/B test). This externally valid "between-subjects" approach models the experience of learning about policy change and experimentation in the real world. When responding to unilateral policy changes—or experiments designed to test them—people rarely learn about foregone alternatives.

Mislavsky et al. (8) found that participants rated low-stakes corporate A/B tests as no worse than their least-preferred policy. These authors also claimed (9) that some of Meyer et al.'s (3) data lacked evidence for experiment aversion. They argued that a proper test of experiment aversion requires comparing each individual's evaluation of their own least-preferred policy to their evaluation of the corresponding A/B test.

This conflict has important implications for research and policy. Objecting to an experiment only because one objects to one or both policies the experiment contains (8, 9) does not necessarily constitute a judgment anomaly. If that were the only reason why people object to experiments, then policy makers could theoretically forestall backlashes to A/B testing by only comparing policies that people like, as Dietvorst et al. (10) suggest. But, if people sometimes object to A/B tests more than they object to either of the policies these tests compare, and absent any rational reasons that might exist for objecting to particular experiments, such a pattern may threaten evidence-based practices and policy by reflecting a genuine aversion to randomized evaluation (3, 11).

We resolve this conflict via five preregistered experiments in which participants evaluate all three options: policy A, policy B, and their A/B test. This within-subjects design allows us to definitively test whether people object to A/B tests more so than to the policies the A/B tests contain. These experiments also test the possibility—suggested by research on joint versus separate evaluation (12–14)—that providing more information about the available options may reduce resistance to randomized evaluation. Learning that a decision-maker chose policy A, and explicitly chose not to test its effectiveness, may reduce the A/B Effect by improving reactions to A/B testing (or by decreasing approval of untested policies).

We used Mislavsky et al.'s (8, 9) preferred statistical criteria and three scenarios they preferred from Meyer et al. (3)—consumer genetic testing, retirement savings options, and autonomous vehicle design. We also tested two of the most-studied and important domains from Meyer et al.: hospital safety checklists and comparative drug effectiveness. Participants rated the appropriateness of and rank-ordered each of three alternative decisions available to a leader: implement policy A, implement policy B, or conduct an A/B test to learn which policy is more effective and implement it for everyone going forward.

Results

In all five experiments, more participants objected to A/B tests (by rating them somewhat or very inappropriate) than objected to either policy (Fig. 1, *Top*). Participants also demonstrated the A/B Effect by rating each A/B test as less appropriate than their average rating of policies A and B (mean [A,B]; $ps \leq 0.01$; Table 1 reports inferential statistics and effect sizes). In all but one experiment ("Autonomous Vehicles"), participants met Mislavsky et al.'s (8, 9) more stringent criterion for "experiment aversion" by rating the A/B test as less appropriate than their least-preferred policy (min [A,B]) ($ps < 0.001$; Fig. 1 *Bottom*). In four of five scenarios, including two of the three Mislavsky et al. preferred, participants therefore viewed unilateral implementation of untested policies as more appropriate than an A/B test designed to evaluate these policies—even when the policy being implemented was their least-favorite of the two.

Across experiments, 50% of participants rated the A/B test as less appropriate than their average policy rating (ranging from 40 to 53% across experiments; Table 1). Over one third of participants (37%; 22 to 46% across experiments) rated the A/B test as less appropriate than either policy. Nearly one-quarter (24%; 16 to 27% across experiments) passed a very conservative test by explicitly objecting to the A/B test while not objecting to either policy (rating neither as inappropriate). Even in the scenario that failed the "experiment aversion" test (8, 9), a nontrivial percentage—16%—of participants met this conservative criterion.

The most popular choice was to rank the A/B test as the worst option (44% across experiments; ranging from 31 to 55%).

Author contributions: P.R.H., C.F.C., D.J.W., and M.N.M. designed research; P.R.H. performed research; P.R.H. analyzed data; P.R.H., C.F.C., and M.N.M. wrote the paper; P.R.H. prepared preregistrations and data archive; and D.J.W. critically edited the paper.

[1]To whom correspondence may be addressed. Email: pheck1000@gmail.com.

First published July 27, 2020.

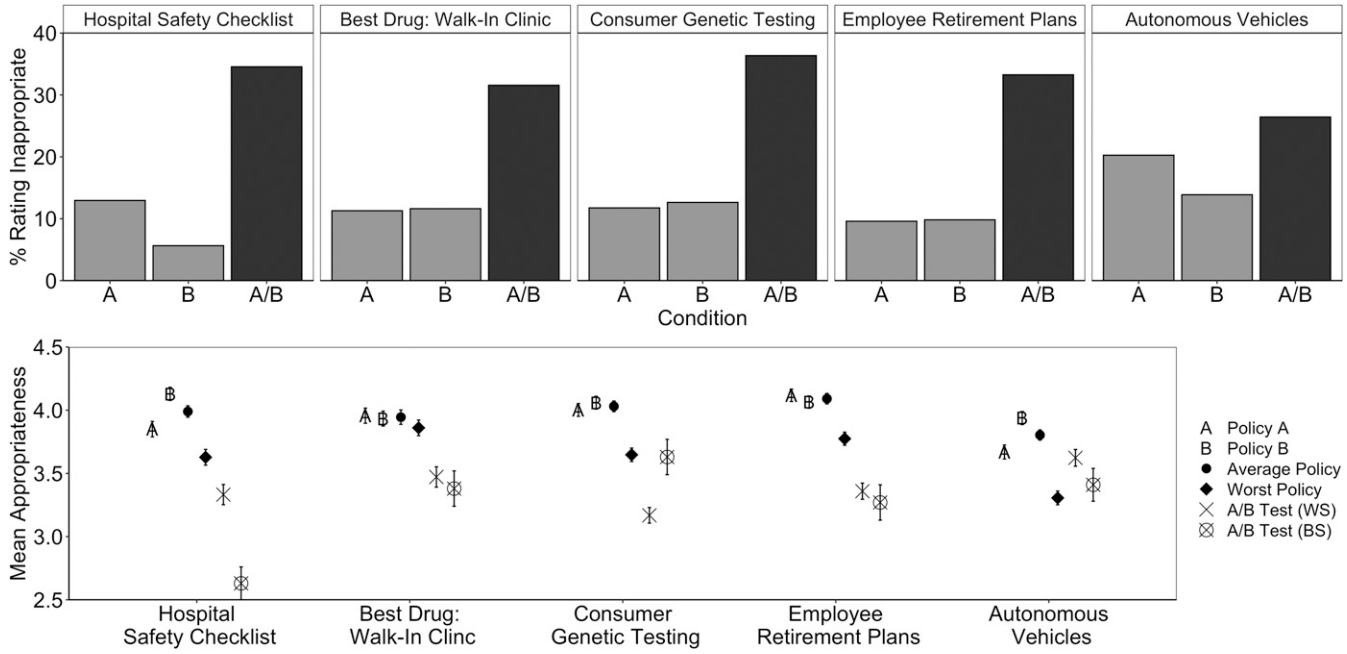

Fig. 1. (*Top*) Percentages of participants objecting to implementing policy A, policy B, or running an A/B test (experiment). (*Bottom*) Mean appropriateness ratings, with SEs, for the A, B, and A/B conditions. "A/B Test (WS)" refers to the A/B condition in the present studies using a within-subjects design; "A/B Test (BS)" refers to previous A/B condition ratings from a between-subjects design (3).

However, a substantial minority ranked the A/B test as the best option (37% across experiments; ranging from 23 to 59%).

We observed no consistently significant associations between participant demographics and ratings of the A/B test condition.

Discussion

Across five preregistered experiments with 1,955 participants, we found that people can object to A/B tests despite approving of unilateral implementation of both untested policies, and despite having information about the alternative options that a decision-maker could have chosen. In four out of five domains, people tended to prefer direct implementation to rigorous evaluation of untested policies, even when they judged one policy to be superior to the other. Converging measures and tests refuted the hypothesis that people object to A/B tests only when they contain a policy the rater finds undesirable (8–10).

People can rationally object to experiments, or decline to participate in them, for a variety of reasons, such as when one treatment is known to be superior or when the two treatments involve preference sensitive trade-offs (3). We are specifically interested in cases where these rational objections do not pertain. In the comparative drug effectiveness scenario (Best Drug: Walk-In), for instance, patients in both the policy and the A/B conditions randomly receive different (and perhaps unequal) treatments, apparently without providing consent. This is because, in walk-in clinics and emergency rooms, patients see whichever doctor happens to be available, and doctors vary in which approved drugs they prescribe for reasons having nothing to do with evidence of efficacy (14). Yet we still often find substantial A/B Effects.

The new measure of rank-order preference also revealed experiment aversion. However, many participants (a majority in one domain and a plurality in another; see Table 1) ranked the A/B test condition as the best option. It is unclear why we did not observe experiment aversion in the Autonomous Vehicles scenario. These results suggest that randomized evaluation can polarize attitudes and that some people prefer experimentation in certain cases. Uncovering heterogeneity in the effect across distinct measures and populations may help policy makers learn when and how best they can reduce costly objections to randomized evaluation.

Reactions to randomized experiments may vary with the amount and type of accompanying information (12, 13), and with the amount of uncertainty regarding treatment effects that these experiments are expected to resolve. When people judge an A/B test without being alerted to the possibility of giving A or B to everyone (e.g., the studies in ref. 3), their ratings may differ from ratings of these same experiments by individuals who learn that it was possible to simply apply A or B to everyone (e.g., the studies reported here). Indeed, ratings of the A/B test were sometimes higher, sometimes lower, and sometimes equivalent when they were elicited using a between- versus within-subjects design (Fig. 1, *Bottom* compares these results with ref. 3). If the information available affects reactions to A/B tests, then providing more (or less) information, and paying careful attention to how the policy or experiment is described, may improve reactions to randomized experiments. Organizations may be able to reduce or eliminate the A/B Effect by framing A/B testing as a superior alternative to universally implementing untested policies, or by describing these policies as what they often are: a shot in the dark based on the highest-paid person's opinion.

Methods

Participants. Preregistered sample sizes [n = ~300 for scenarios with large A/B Effects and n = ~450 for scenarios with small-to-medium A/B Effects (3)] were chosen for 95% power to detect d = 0.21 and d = 0.17, respectively (two-tailed paired t test, $\alpha = 0.05$). MTurk participants received $0.40 and were excluded if they participated in our other studies on this topic.

Materials and Procedure. We adapted five scenarios (3) to create within-subjects designs: "Hospital Safety Checklist," "Best Drug: Walk-In Clinic," "Consumer Genetic Testing," "Employee Retirement Plans," and "Autonomous Vehicles." Participants read all three conditions (A, B, and A/B), which were presented on the same page in counterbalanced order, and then rated the appropriateness of each decision on a 1 to 5 scale on the same page and in

Table 1. Descriptive and inferential statistics for tests of the A/B Effect and "experiment aversion"

	Descriptive results				Inferential results	
Scenario	Variable	Mean (SD) rating	Rank: Best	Rank: Worst	Test description	Test outcome
Hospital Safety Checklist (n = 301)	A	3.85 (1.06)	26%	31%	Mean(A,B) vs. Mean(AB)	t = 7.53***, d = 0.58 ± 0.16
	B	4.13 (0.90)	38%	24%	Min(A,B) vs. Mean(AB)	t = 3.23**, d = 0.24 ± 0.15
	AB	3.33 (1.39)	37%	46%	Mean(A,B) < AB	53%*** ± 6%
	Mean(A,B)	3.99 (0.78)			Min(A,B) < AB	37%*** ± 6%
	Min(A,B)	3.63 (1.08)			AB = 1,2 & A,B = 3,4,5	27%*** ± 5%
Best Drug: Walk-In Clinic (n = 301)	A	3.96 (1.04)	22%	29%	Mean(A,B) vs. Mean(AB)	t = 4.77***, d = 0.39 ± 0.17
	B	3.93 (1.02)	19%	35%	Min(A,B) vs. Mean(AB)	t = 3.85***, d = 0.31 ± 0.16
	AB	3.47 (1.40)	59%	37%	Mean(A,B) < AB	43%*** ± 6%
	Mean(A,B)	3.95 (0.99)			Min(A,B) < AB	40%*** ± 6%
	Min(A,B)	3.86 (1.08)			AB = 1,2 & A,B = 3,4,5	27%*** ± 5%
Consumer Genetic Testing (n = 451)	A	4.00 (1.07)	34%	26%	Mean(A,B) vs. Mean(AB)	t = 12.99***, d = 0.76 ± 0.13
	B	4.06 (1.08)	43%	19%	Min(A,B) vs. Mean(AB)	t = 6.58***, d = 0.39 ± 0.12
	AB	3.17 (1.31)	23%	55%	Mean(A,B) < AB	59%*** ± 5%
	Mean(A,B)	4.03 (0.89)			Min(A,B) < AB	46%*** ± 5%
	Min(A,B)	3.65 (1.13)			AB = 1,2 & A,B = 3,4,5	27%*** ± 5%
Employee Retirement Plans (n = 448)	A	4.12 (1.04)	37%	27%	Mean(A,B) vs. Mean(AB)	t = 10.29***, d = 0.64 ± 0.13
	B	4.06 (1.02)	29%	25%	Min(A,B) vs. Mean(AB)	t = 5.46***, d = 0.34 ± 0.12
	AB	3.36 (1.35)	34%	49%	Mean(A,B) < AB	53%*** ± 5%
	Mean(A,B)	4.09 (0.88)			Min(A,B) < AB	42%*** ± 5%
	Min(A,B)	3.77 (1.10)			AB = 1,2 & A,B = 3,4,5	26%*** ± 4%
Autonomous Vehicles (n = 454)	A	3.67 (1.18)	24%	44%	Mean(A,B) vs. Mean(AB)	t = 2.52*, d = 0.15 ± 0.12
	B	3.94 (1.10)	34%	26%	Min(A,B) vs. Mean(AB)	t = –4.28, d = –0.25 ± 0.11
	AB	3.62 (1.41)	42%	31%	Mean(A,B) < AB	40%*** ± 4%
	Mean(A,B)	3.80 (0.87)			Min(A,B) < AB	22%*** ± 4%
	Min(A,B)	3.31 (1.16)			AB = 1,2 & A,B = 3,4,5	16%*** ± 3%

The "Scenario" column lists vignettes and sample sizes for studies 1–5. The next four columns display each study's descriptive results. The last two columns report five hypothesis tests for each study, each assessing a different criterion for the A/B Effect (ABE). The first test evaluates the original criterion (3) and was always preregistered as confirmatory. The second test evaluates Mislavsky et al.'s proposed criterion (9). The remaining tests compare the observed percentage of participants meeting the stated criterion against a null hypothesis of zero, which would indicate no "experiment aversion." For the fifth test, "AB = 1,2" indicates a rating of very or somewhat inappropriate, and "AB = 3,4,5" indicates a rating that is not explicitly inappropriate. Symbols denote statistical significance for ABE: ***P < 0.001; **P = 0.001, *P = 0.01.

the same order as the decisions were presented. Participants then rank-ordered the decisions, explained their responses in a text box, and provided demographics.

These anonymous online studies were determined to be exempt by the Geisinger Internal Review Board, and informed consent was not required.

Data Availability. Participant response data, preregistrations, materials, and analysis code have been deposited in Open Science Framework (https://osf.io/w6qub/) (15).

ACKNOWLEDGMENTS. We thank Pedram Heydari and Anh Huynh for their contributions.

1. D. Baldassarri, M. Abascal, Field experiments across the social sciences. *Annu. Rev. Sociol.* **43**, 41–73 (2017).
2. J. D. Lantos, Randomized trials are deeply offensive. *Am. J. Bioeth.* **20**, 3–5 (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. S. Hellman, D. S. Hellman, Of mice but not men. Problems of the randomized clinical trial. *N. Engl. J. Med.* **324**, 1585–1589 (1991).
5. M. N. Meyer, Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colo. Tech. L. J.* **13**, 273–331 (2015).
6. H. F. Lynch, D. J. Greiner, I. G. Cohen, Overcoming obstacles to experiments in legal practice. *Science* **367**, 1078–1080 (2020).
7. K. Thomas, Trump calls this drug a "game changer." Doctors aren't so sure. *The New York Times*, 18 April 2020, Section A, p. 15.
8. 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 (2019).
9. 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).
10. B. Dietvorst, R. Mislavsky, U. Simonsohn, Experimentation aversion: Reconciling the evidence. *Data Colada* (2019). http://datacolada.org/79. Accessed 7 November 2019.
11. 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).
12. M. H. Bazerman, D. A. Moore, A. E. Tenbrunsel, K. A. Wade-Benzoni, S. Blount, Explaining how preferences change across joint versus separate evaluation. *J. Econ. Behav. Organ.* **39**, 41–58 (1999).
13. C. K. Hsee, The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organ. Behav. Hum. Decis. Process.* **67**, (1996).
14. J. E. Wennberg, Dealing with medical practice variations: A proposal for action. *Health Aff. (Millwood)* **3**, 6–32 (1984).
15. P. R. Heck, C. F. Chabris, D. J. Watts, M. N. Meyer, Online archive for Objecting to Experiments Even While Approving of the Policies or Treatments They Compare. Open Science Framework. https://osf.io/w6qub. Deposited 2 July 2020.