Objecting to experiments that compare two unobjectionable policies or treatments

Michelle N. Meyer[a,1], Patrick R. Heck[a,b,2], Geoffrey S. Holtzman[a,b,2,3], Stephen M. Anderson[a,b,4], William Cai[c,5], Duncan J. Watts[c], and Christopher F. Chabris[b,d]

[a]Center for Translational Bioethics and Health Care Policy, Geisinger Health System, Danville, PA 17821; [b]Autism and Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA 17837; [c]New York City Lab, Microsoft Research, New York, NY 10011; and [d]Institute for Advanced Study in Toulouse, 31015 Toulouse, France

Edited by Dalton Conley, Princeton University, Princeton, NJ, and approved April 8, 2019 (received for review December 5, 2018)

Randomized experiments have enormous potential to improve human welfare in many domains, including healthcare, education, finance, and public policy. However, such "A/B tests" are often criticized on ethical grounds even as similar, untested interventions are implemented without objection. We find robust evidence across 16 studies of 5,873 participants from three diverse populations spanning nine domains—from healthcare to autonomous vehicle design to poverty reduction—that people frequently rate A/B tests designed to establish the comparative effectiveness of two policies or treatments as inappropriate even when universally implementing either A or B, untested, is seen as appropriate. This "A/B effect" is as strong among those with higher educational attainment and science literacy and among relevant professionals. It persists even when there is no reason to prefer A to B and even when recipients are treated unequally and randomly in all conditions (A, B, and A/B). Several remaining explanations for the effect—a belief that consent is required to impose a policy on half of a population but not on the entire population; an aversion to controlled but not to uncontrolled experiments; and a proxy form of the illusion of knowledge (according to which randomized evaluations are unnecessary because experts already do or should know "what works")—appear to contribute to the effect, but none dominates or fully accounts for it. We conclude that rigorously evaluating policies or treatments via pragmatic randomized trials may provoke greater objection than simply implementing those same policies or treatments untested.

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

Randomized experiments, also known as randomized controlled trials (RCTs) or A/B tests, have long been the "gold standard" for evaluating drugs and other medical interventions and are increasingly used to evaluate business products and services, government programs, education and health policies, and global aid (1–6). Despite their critical role in advancing human welfare, randomized experiments raise legitimate ethical concerns. For example, if an experiment randomizes some participants to a treatment that is already known to be inferior to the standard of care available outside of the trial, many ethicists would deem such a trial unethical. Even when the relevant expert community is uncertain as to which of two treatments is more effective, individuals may prefer the side effects of one treatment to the other (as was likely the case in the historic RCT comparing mastectomy with lumpectomy) (7). In such cases, many of us reasonably decline to sacrifice our own welfare or autonomy by participating in particular trials, even if the knowledge produced by the trial is expected to help others.

But do we also object to randomized experiments even when (*i*) neither treatment is known to be inferior and (*ii*) we would not object to either treatment if we received it deterministically (i.e., if everyone received it)? Anecdotal evidence suggests that we may. For example, Pearson Education came under public criticism after the media reported that it had randomized math and computer science students at different schools to receive one of three versions of its instructional software: two versions displayed different encouraging messages as students attempted to solve problems, while a third displayed no messages (8). To our knowledge, no one had objected to the previous software, which—because of a unilateral choice made by the company—provided no encouragement. Had Pearson Education instead chosen to display encouraging messages to all students in its software, it is unlikely that any users would have objected. However, briefly randomizing different schools to receive each of these individually unobjectionable conditions was condemned. Experiments by Facebook to determine whether positive and/or negative posts negatively impact users' happiness, by OkCupid (a dating website) to compare the effectiveness of its matching algorithm to that of the power of suggestion, by physicians to compare treatment options for premature babies within the existing standard of care, and by medical residency programs to determine whether more or less

Significance

Randomized experiments—long the gold standard in medicine—are increasingly used throughout the social sciences and professions to evaluate business products and services, government programs, education and health policies, and global aid. We find robust evidence—across 16 studies of 5,873 participants from three populations spanning nine domains—that people often approve of untested policies or treatments (A or B) being universally implemented but disapprove of randomized experiments (A/B tests) to determine which of those policies or treatments is superior. This effect persists even when there is no reason to prefer A to B and even when recipients are treated unequally and randomly in all conditions (A, B, and A/B). This experimentation aversion may be an important barrier to evidence-based practice.

Author contributions: M.N.M., P.R.H., G.S.H., D.J.W., and C.F.C. designed research; M.N.M., P.R.H., G.S.H., S.M.A., W.C., D.J.W., and C.F.C. performed research; M.N.M., P.R.H., G.S.H., S.M.A., W.C., D.J.W., and C.F.C. analyzed data; and M.N.M., P.R.H., G.S.H., D.J.W., and C.F.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Preregistrations, data, and analysis scripts related to this study have been deposited in the Open Science Framework, https://osf.io/5y4f9/.

[1]To whom correspondence should be addressed. Email: michellenmeyer@gmail.com.

[2]P.R.H. and G.S.H. contributed equally to this work.

[3]Present address: Department of Psychology, Franklin & Marshall College, Lancaster, PA 17603.

[4]Present address: Department of Psychology, The Pennsylvania State University, University Park, PA 16802.

[5]Present address: Department of Management Science and Engineering, Stanford University, Stanford, CA 94305.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820701116/-/DCSupplemental.

Published online May 9, 2019.

flexible working hours result in better patient outcomes all generated similar controversy among a wide variety of lay and expert commenters, including members of the media, members of the general public, ethicists and other academics, advocacy organizations, and government officials and lawmakers (9–13).

Although each of these controversies likely reflects a unique combination of concerns, we propose that they all exhibit a common pattern that has previously been labeled the "A/B illusion" (9): people appeared to judge a randomized experiment comparing two unobjectionable policies or treatments (A and B), neither of which was known to be superior, as less appropriate than simply implementing either A or B for everyone. Although in many of these cases meaningful consent to the A/B test was lacking, it was—crucially—equally lacking when the policies were imposed universally. Under such conditions, randomly assigning half of people to A and the other half to B did not impose an unacceptable policy on anyone and was no different with respect to individual autonomy than imposing either A or B on everyone. Objecting to the A/B test but not to universal imposition of either policy condition is therefore puzzling, and also potentially problematic for advancing understanding of policy effectiveness (9). For example, in the Pearson Education case, the A/B test showed that students who received no encouragement actually attempted more problems. Had Pearson Education anticipated the public reaction to its A/B test, instead of conducting a randomized experiment, the company may well have followed its instincts and implemented an inferior version of the software for everyone.

Unfortunately, evidence from public controversies is highly selective: We only know of explicit negative judgments about A/B tests from those who are motivated to voice their opinion (e.g., on Twitter or in the media), and we infer from the comparative silence that people are neutral toward the unilateral imposition of A or B policies. Here we sought to determine whether we could systematically observe this effect in a variety of domains where field experiments are commonly employed and to evaluate potential explanations for it. We conducted a series of online vignette studies [all but the first of which were preregistered at Open Science Framework (OSF)] with US residents via Amazon Mechanical Turk (MTurk) and Pollfish and with healthcare professionals (total n = 5,873). We randomly assigned participants to rate the appropriateness of a fictitious agent's decision to implement one policy (the A condition), implement another policy (the B condition), or conduct a randomized experiment comparing A and B (the A/B condition). We also asked participants to briefly explain their ratings (which we qualitatively coded in four experiments) and collected demographic information and a measure of scientific literacy. Detailed statistical results for all studies are provided in *SI Appendix*.

In all studies, we chose pairs of policies that participants (often in pretesting) judged to be roughly equivalent in appropriateness and which they judged to be appropriate overall (i.e., above the scale midpoint). Beginning with the second of our 16 studies, we preregistered our methods and hypotheses, including the prediction that participants would object more to an A/B test that compared two unobjectionable policies than to the implementation of either policy alone. Importantly, all vignettes were silent in both the policy and A/B conditions about whether the agent planned to seek the consent of the affected parties before initiating a new policy or an A/B test of those policies. The reason is that participants in either the A-only condition or the B-only condition are in the same position as the A-treatment and B-treatment participants in the randomized condition. If respondents in the A/B condition demand that informed consent be obtained, then a consistent response would be to demand the same in the A and B conditions. We preregistered the hypothesis, however, that respondents would inconsistently object to the apparent lack of consent, doing so at a significant rate in the A/B conditions, but not in the policy conditions.

Results

In study 1 (n = 413 MTurk participants), participants read a short description of a hospital director who wants to reduce deadly and costly catheter-related hospital infections and thinks that providing doctors with a checklist of standard safety precautions might help [note: studies of this sort are often classified as "quality improvement" projects and therefore not subject to federal regulations governing "research," including review by an institutional review board (IRB) and consent]. We designed our stimuli to depict realistic communication of decisions within organizations. A/B tests were described, as they often are, as intended to improve a situation by determining which of two options is superior. Policy implementations were described, as they typically are, as fiats from management that make no mention of any evidence base to support them. Our participants were then randomly assigned to read about one of four decisions and rate its appropriateness on a 1–5 Likert scale (1 = very inappropriate; 3 = neither inappropriate nor appropriate; 5 = very appropriate):

> Badge (A): The director decides that all doctors who perform this procedure will have the standard safety precautions printed on the back of their hospital ID badges.

> Poster (B): The director decides that all rooms where this procedure is done will have a poster displaying the standard safety precautions.

> A/B short: The director decides to run an experiment by randomly assigning patients to be treated by a doctor wearing the badge or in a room with the poster.

> A/B learn: Same as A/B short, with an added sentence noting that after a year, the director will have all patients treated in whichever way turns out to have the highest survival rate.

Here and in several other studies, we included two A/B conditions but observed no significant differences between them (see below). We eventually discontinued A/B short and, for studies in which we ran both, combined them for all analyses reported here.

Fig. 1*A* shows the percentage of people in the three collapsed conditions who rated the decision as inappropriate (responses of 1 or 2 on the five-point scale). Treating the ratings as continuous variables, the badge (A) was rated as more appropriate (M = 3.93, SD = 1.14) than the A/B tests [M = 2.74, SD = 1.31; $t(293) = 7.01$, $P < 0.001$, $d = 0.94$, 95% CI: (0.66, 1.21)] and the poster (B) was also rated as more appropriate (M = 4.35, SD = 1.0) than the A/B tests [$t(336) = 11.58$, $P < 0.001$, $d = 1.32$, 95% CI: (1.08, 1.57)]. Here, and throughout all experiments, the pattern of results was similar for percentage and continuous variable formats (see *SI Appendix* for both).

Study 2 replicated the results of study 1 on MTurk using identical materials (n = 386; Fig. 1*B*) and using longer versions of the vignettes (n = 343; Fig. 1*C*), and with a sample of mobile device users recruited by Pollfish using the study 1 materials (n = 679; Fig. 1*D*). In these preregistered replications, the A and B conditions were each judged significantly more appropriate than both A/B conditions, all $P < 0.001$, though effect sizes were smaller than in study 1.

Perhaps the effect is limited to the medical domain, where the doctor–patient relationship is somewhat unique. To find out, in study 3 (n = 2,270), we investigated several nonmedical domains: direct-to-consumer genetic testing, autonomous vehicle design, employee retirement plan enrollment nudges, recruitment of health workers in developing countries, alleviation of extreme poverty, promoting school teacher well-being, and basic income policy options. We observed significant effects in the first six domains, with an average effect size of d = 0.44, but not for the seventh (d = 0.10) (Table 1; see *SI Appendix, Robustness Checks on*

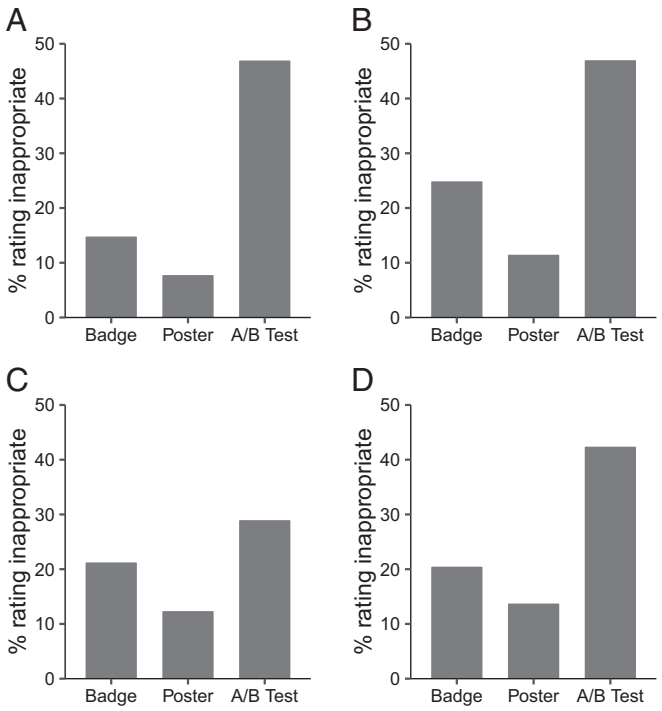

Fig. 1. Results of safety checklist study and replications (studies 1 and 2). (*A*) Initial MTurk experiment; (*B*) direct replication; (*C*) replication with alternate vignette; (*D*) replication on Pollfish platform. Responses were made on a five-point scale but are presented here as percentage of participants who chose "very inappropriate" or "somewhat inappropriate," to reflect the rate of disapproval.

Selection Bias and Multiple Comparisons for additional analyses performed using *p* curve, Bonferroni correction, and hierarchical linear modeling to guard against possible selection bias and correct for multiple comparisons across studies).

The results of studies 1–3 suggest that we have identified an A/B effect that is both robust and general. But why does it occur? To explore seven possible causal mechanisms of our effect, we conducted three additional studies, investigated whether demographic factors explained any variance in appropriateness ratings, and examined participants' free response comments in four experiments.

First, participants may have rated the A and B conditions as generally appropriate because they failed to imagine superior alternatives. In that case, the relatively lower appropriateness ratings in the A/B condition might simply reflect participants engaged in joint evaluation of the treatments (14) who strongly prefer one policy to object to randomly assigning people to treatments they perceive (whether correctly or not) to be unequal. For instance, people might independently judge both mastectomy and lumpectomy to be unobjectionable treatments for breast cancer, but when jointly confronted with both options in the form of an A/B test comparing the two, they may (rightly) object that some women will prefer one option to the other, even if current evidence suggests that both are otherwise equally effective in treating cancer. Failure to imagine other, potentially superior, alternatives frequently occurs when agents implement policies, as do (often evidence-free) intuitions that one treatment in an A/B test is superior. Both might explain why people tend not to object to universal application of untested policies but do object to A/B tests of these policies. Indeed, 14% of participants in the A/B conditions of study 1 and its first replication commented that one policy was preferable or that the A/B test treated people unequally. Notably, excluding these participants still yields a substantial A/B effect in study 1, $t(388) = 11.3$, $P = 0.001$, $d = 1.14$, and its replication (study 2a), $t(354) = 7.30$, $P = 0.001$, $d = 0.78$ (see *SI Appendix*, Tables S6 and S12 for this analysis).

To further determine whether the effect remains when neither a failure of imagination nor a rational preference for one policy over the other can plausibly be involved, we conducted two additional MTurk studies pertaining to another important domain: comparative effectiveness drug trials. Many people assume their healthcare providers choose their medications and other clinical interventions based on scientific evidence. But the Food and Drug Administration (FDA) approval process leaves open many important questions about drugs' effects in real-world conditions, for off-label purposes, and compared with other drugs. In the absence of evidence about such effects, unjustified "variation in medical practices" arising out of subjective attitudes or idiosyncratic experiences of individual physicians is ubiquitous (15). One important solution is so-called phase IV postmarketing trials. Unlike clinical trials of novel drugs, these and other minimal-risk pragmatic trials do not always require informed consent as a matter of either law or ethics (16–18).

In study 4, participants were told that there exist multiple FDA-approved blood pressure drugs and that, of these, "Doctor Jones" decides to prescribe all his patients a drug named simply "drug A" or "drug B" (in the A and B conditions), or decides to randomly assign his patients to receive drug A or drug B (in the A/B condition). Here, there is no reason for participants jointly evaluating the two policies in the A/B condition to think that A or B is better, and accordingly only 4% of participants raised this concern (Table 2). Nevertheless, we observed an A/B effect, $t(300) = 7.89$, $P < 0.001$, $d = 0.96$. And, as shown in Fig. 2*A*, the percentage of participants objecting to the A/B test is not simply the sum of the percentages objecting to either A or B.

Second, the effect might be explained by an aversion to randomization. Even if respondents do not personally favor either A or B, they may still infer that one treatment must be better, so an experiment in which half receive the "worse" treatment must be unfair. To investigate this possibility, in study 5, we modified the drug scenarios to occur in a walk-in clinic where some doctors prescribe drug A, some prescribe drug B, and "patients see whichever

Table 1. Experiment disapproval observed in multiple domains (study 3)

Scenario	Condition	*N*	% objecting	*M* rating	SD	SEM	A/B effect
Genetic	A	97	15.5	4.13	1.17	0.12	$t(376) = 5.12$
Testing	B	102	8.8	4.30	1.00	0.10	$P < 0.001$
	A/B	179	21.2	3.60	1.26	0.09	$d = 0.53$
Autonomous	A	104	11.5	4.20	1.22	0.12	$t(395) = 4.36$
Vehicles	B	100	19.0	3.98	1.46	0.15	$P < 0.001$
	A/B	193	28.0	3.51	1.33	0.10	$d = 0.44$
Retirement	A	98	19.4	3.76	1.32	0.13	$t(294) = 3.37$
Plans	B	103	19.4	3.90	1.35	0.13	$P < 0.001$
	A/B	95	36.8	3.27	1.32	0.14	$d = 0.42$
Health	A	96	9.4	4.14	0.98	0.10	$t(291) = 3.15$
Worker	B	101	9.9	4.10	1.03	0.10	$P = 0.002$
Recruitment	A/B	96	16.7	3.70	1.19	0.12	$d = 0.39$
Poverty	A	96	24.0	3.57	1.25	0.13	$t(300) = 3.44$
Alleviation	B	103	10.7	4.06	1.15	0.11	$P < 0.001$
	A/B	103	35.0	3.31	1.24	0.12	$d = 0.42$
Teacher	A	99	12.1	4.04	1.10	0.11	$t(298) = 3.34$
Well-being	B	97	27.8	3.64	1.36	0.14	$P < 0.001$
	A/B	104	31.7	3.34	1.25	0.12	$d = 0.41$
Basic	A	102	20.6	3.73	1.15	0.11	$t(302) = 0.73$
Income	B	106	18.9	3.75	1.14	0.11	$P = 0.465$
	A/B	96	21.9	3.64	1.21	0.12	$d = 0.09$

Table 2. Selected coding results for studies 1, 2, 4, and 5: Percentage of participants in each condition who provided each of four reasons for their appropriateness rating

Codes received	Condition			
Study 1 (checklist)	Badge	Poster	A/B learn	A/B
Inequality	0	0	11	11
Consent	0	0	11	7
Experimentation	0	1	34	22
Randomization	0	0	1	3
Study 2a (checklist—direct replication)	Badge	Poster	A/B learn	A/B
Inequality	0	0	17	16
Consent	0	0	7	13
Experimentation	0	0	32	21
Randomization	0	0	8	4
Study 4 (drug effectiveness)	Drug A	Drug B	A/B learn	
Inequality	0	0	4	
Consent	0	0	16	
Experimentation	0	0	18	
Randomization	0	0	6	
Study 5a (drug effectiveness walk-in)	Drug A	Drug B	A/B learn	
Inequality	1	0	2	
Consent	0	2	14	
Experimentation	0	0	21	
Randomization	0	0	4	

doctor is available." Hence, patients in the policy conditions, too, are now effectively randomized to receive drug A or B. Here, the effect was about 30% smaller but still substantial, $t(301) = 5.27$, $P < 0.001$, $d = 0.64$ (Fig. 2*B*). We replicated this finding on Pollfish, where despite a very large effect of response scale order, judgments still followed the pattern we have consistently observed, $t(718) = 1.92$, $P = 0.055$, $d = 0.15$ (Fig. 2*C*). Consistent with these results, no more than 6% of participants in the A/B conditions of any experiment we coded commented negatively about randomization (Table 2).

Third, it could be that people object to the implied absence of informed consent to the A/B tests. As predicted, across all four experiments we coded, 18% of participants in the A/B conditions complained about the apparent lack of either consent by, or notice to, recipients of A/B tests, whereas fewer than 1% of participants in our policy conditions raised the same objection (Table 2). As noted earlier, this inconsistency is a puzzle, as in all conditions, people were subjected without their consent to one of the same two untested policies with unknown effects. Inconsistent beliefs about when consent is ethically required may be an important causal mechanism—or manifestation—of the A/B effect.

Fourth, it could be that people object to experiments they assume are motivated by trivial or nefarious goals. Since at least the mid-19th century, depictions of the "mad scientist" have shaped the public's view of science (19). If this were motivating the A/B effect, however, then the A/B-learn conditions, in which participants were told the purpose and use of the experimental results, should be rated considerably more appropriate than the matched A/B-short conditions. As noted above, we observed almost no difference between these [pooled across 1,302 participants in six experiments: $t(1{,}300) = 1.50$, $P = 0.13$, $d = 0.02$, 95% CI (−0.09, 0.12)].

Fifth, participants may believe that the vignette agents or other experts (e.g., the FDA) either already do or should know what the correct treatment is, a proxy form of illusion of knowledge (20). In the four experiments we coded, 9.5% of participants espoused some version of this belief. Of these 120 participants, the vast majority (74%) cited the proxy illusion of knowledge either to explain why an agent's untested policy implementation was appropriate (56%) or to explain why his A/B test was inappropriate (18%) (*SI Appendix*, Table S44). This pattern makes intuitive sense: raters generally trust experts' judgment of what is best to do, and this trust is undermined when experts acknowledge their uncertainty by contemplating alternative treatments in an A/B test. The proxy illusion of knowledge therefore appears to contribute to the A/B effect in two ways: by making untested policies more acceptable and by making A/B tests less acceptable. The stickiness of the proxy illusion of knowledge is suggested by the fact that participants continue to exhibit it even in study 5. In that vignette, participants in all conditions were told that some doctors prescribe drug A while others prescribe drug B. Participants are given no reason to believe that there was an expertise-based reason for this variation in practice (such as an intuitive form of precision medicine in which each doctor somehow prescribes the best drug for each unique patient before him), because each doctor prescribes a single drug to all of his patients and patients see whichever doctor is available when they walk into the clinic. However, participants in the policy condition of study 5 invoked the proxy illusion of knowledge to support the expert's universal implementation more often than did participants in the policy conditions of our other three experiments. As with the other mechanisms contributing to the A/B effect, however, the proxy illusion of knowledge appears to explain only a small portion of the effect, as relatively few participants (9.5%) articulated this view.

Sixth, people might be averse to experimentation per se. In the four experiments we coded, 24% of participants in the A/B

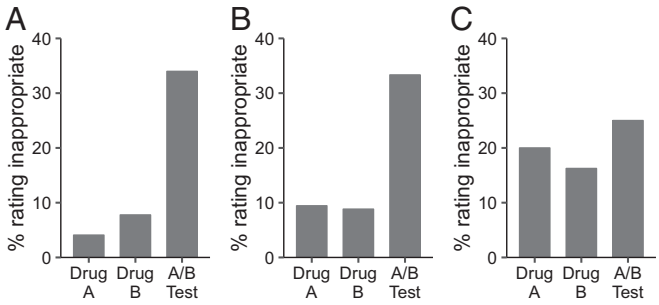

Fig. 2. Disapproval of experiments not explained primarily by joint evaluation or aversion to randomization. (*A*) study 4, MTurk; (*B*) study 5, MTurk; (*C*) study 5, Pollfish.

conditions described the vignette agent as "playing with lives," treating people like "guinea pigs," or otherwise inappropriately "experimenting." But, as with our finding about consent, it is striking that only a single individual out of 791 participants complained about the "experimental" nature of the policy conditions, which involve the same risks and uncertainties as the A/B tests—and hence equally entail "gambling with lives"—but are simply uncontrolled. It could be that participants have strong negative associations with the language of "experiments" (21), which we used to describe the A/B tests but not the policies.

Finally, it could be that people object to experimentation when they are relatively less educated, relatively less educated in the sciences, have comparatively poor scientific literacy, or have less expertise in the domain in which A/B testing or policy implementation is proposed. To investigate these possibilities, we first examined whether educational attainment, having a science, technology, engineering, and mathematics (STEM) degree, or scientific understanding explained any of the variance in appropriateness ratings of experiments. They (and the other demographic variables we collected) explain almost no variance (*SI Appendix*, Fig. S20 and Tables S55–S64). Next, we conducted study 6, in which we replicated study 1 and study 5—our scenarios involving medicine and healthcare delivery—in a sample of healthcare providers employed by Geisinger, a large US health system. Prior research with Geisinger employees sought to determine support and preparedness for establishing Geisinger as a "learning heath system" (4), in which research is seamlessly embedded into practice to enable continuous learning and improvement. A survey found that 98% ($n = 126$) of respondents (most of whom were clinicians) agreed that "evidence supports the claim that a learning health system is necessary to provide safe, effective, and beneficial patient-centered care at lower cost," with 53% of the sample strongly agreeing with this statement (22). Interviews ($n = 41$) with Geisinger leadership similarly found unanimous support for "the general concept and goals" of the learning healthcare system and for "enhancing learning across the institution" (23). However, in study 6, in both the safety checklist, $t(224) = 6.09$, $P < 0.001$, $d = 0.86$ (Fig. 3*A*), and drug effectiveness scenarios, $t(229) = 6.26$, $P < 0.001$, $d = 0.87$ (Fig. 3*B*), we found the same pattern as in our previous studies with laypersons, with comparable effect sizes.

Discussion

We find evidence across 16 studies of 5,873 participants from three populations spanning nine domains—from healthcare to autonomous vehicle design to policies to address global poverty—that people frequently rate field experiments designed to establish comparative effectiveness of two policies as inappropriate even when the policies those experiments compare are widely seen as appropriate. This A/B effect remains robust after a variety of procedures to correct for multiple comparisons, including *p* curve, Bonferroni correction, and hierarchical linear modeling.

The effect persists even when there is no reason to prefer policy A to policy B and even when recipients are already being treated both unequally and randomly in the policy conditions. Several remaining explanations—a belief that consent is required to impose a policy on half of a population but not on the entire population; an aversion to controlled but not to uncontrolled experiments; and the proxy illusion of knowledge—appear to contribute to the effect and should be further explored, but none dominates or fully accounts for it.

Additionally, the effect is just as strong among those with higher educational attainment and science literacy and those with STEM degrees, and among professionals in the relevant domain. Although laypersons generally do not decide whether policies will first be randomly evaluated or immediately implemented, untested, their attitudes toward A/B tests compared with universal implementation nevertheless matter. Policymakers who perceive that recipients will object to randomized evaluations may forgo them in favor of universal implementation or may conduct randomized evaluations in secret, neither of which is optimal (10). Still, it clearly also matters whether those who are in a position to implement untested policies and practices themselves tend to object to experiments comparing two unobjectionable options. The results of our final study, of healthcare providers, suggests that they do.

To be sure, not every ethical objection to an A/B test—even one that compares two accepted practices—involves a logical inconsistency. Moreover, it is often perfectly reasonable to insist that we consent before untested policies or practices are imposed on us. But when neither of two policies is objectionable or perceived as clearly superior, an A/B test comparing them should not be seen as more morally problematic than a unilateral decision to implement either untested policy. The fact that participants consistently—albeit not always—have exactly this reaction across a variety of domains and under a variety of conditions suggests that many real-world ethical objections to randomized experiments also reflect a general pattern in which we judge a formal controlled experiment to be uniquely morally problematic, when in fact the rest of the world outside the experiment is often just the A condition of an A/B test that was never conducted. Indeed, our vignettes presented experiments neutrally or even positively, emphasizing what is to be learned, whereas media accounts often describe them using inflammatory language (11). If anything, therefore, our results may underestimate the degree to which people object to A/B tests in the real world.

The A/B effect explored here may be one of many factors that explain why it is difficult to scale the results of small laboratory or field experiments up to the level of larger populations (24). In particular, a tendency to avoid conducting appropriate experiments in the first place (in favor of implementing untested policies and treatments) could cause policymakers to resist running large-scale field experiments, either because they themselves do not like the experiments or because they fear their constituents and/or other stakeholders will react negatively to learning they were "experimented on." Additionally, people who do not like experiments or do not think they are necessary or useful may be unlikely to volunteer for and more likely to drop out of experiments, contributing to an "adverse heterogeneity bias" (24) by making the participant pool of small-scale initial experiments more homogeneous and thus less representative of the larger-scale population to which the results are meant to generalize. Conversely, those who do volunteer for small initial experiments

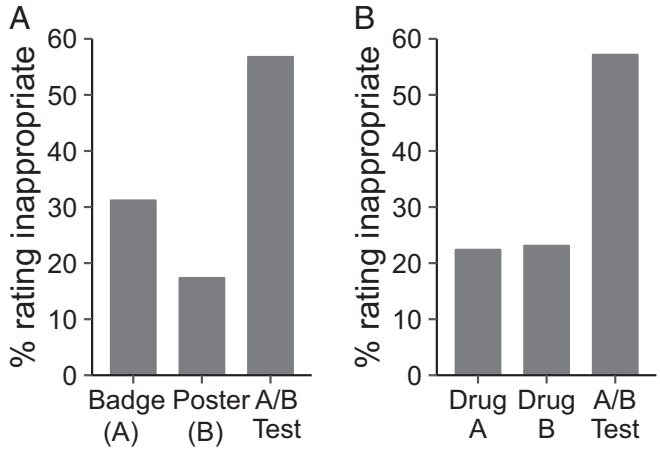

Fig. 3. Results of safety checklist and drug effectiveness replications in healthcare clinicians (study 6). (*A*) safety checklist; (*B*) drug effectiveness.

may be more enamored of researchers and of research (and they may tend to exhibit a small-to-zero A/B effect) and thus more likely to produce results consistent with study hypotheses—which normally predict large, positive effects of interventions that later shrink considerably or even disappear at larger scales.

Disapproving reactions to experimentation may partly reflect a lack of familiarity. The first true randomized experiment was conducted less than two centuries ago (25), and the statistical theory underlying experiments dates only to the 1920s (26, 27). A/B tests are therefore not an intuitive, common-sense, evidence-creating mechanism, and they may simply be hard for us to think clearly about. The A/B effect may also reflect a heuristic about the ethics of experiments (e.g., as distinctively risky or uncertain) that often serves us well but sometimes leads us astray, ironically resulting in patients and other recipients of untested policies and practices being subjected to more risks than they would had A/B tests been conducted first (28, 29). More research is needed to investigate the effect's causal mechanisms and its boundary conditions; in particular, to see (*i*) whether the effect occurs in the context of other forms of experimentation, such as natural experiments and nonexperimental studies in which causation is inferred using sophisticated observational methods, and (*ii*) whether the effect varies in size or presence across different domains (e.g., business, healthcare, social policy, education). Further investigation should also determine the effect's causal mechanisms and develop and test debiasing strategies. Regardless of the reasons, the unfortunate lesson for those who care about evidence-based practice is that implementation of an untested policy based on intuition about what works may be less likely to invite objection than rigorous evaluation of two or more otherwise unobjectionable policies.

Materials and Methods

Participants. All experiments and replications (n = 6,141; n = 5,873 after exclusions of repeat participants) were determined to be exempt from review by the IRB at Geisinger, as were all pretest and pilot studies (n = 1,405 participants; because one pretest was run with repeated measures, n = 2,137 vignette responses). All participants were recruited via MTurk and were paid standard fees, except for the experiments in studies 2 and 5 conducted via Pollfish, and for study 6, where participants were healthcare providers working at Geisinger recruited by email.

Study Format. Each study presented a short vignette in text form and asked for a single rating of a decision's appropriateness on a 1–5 scale, followed by a free-text explanation of why the participant gave the rating they did, followed by a series of demographic and other questions. In all experiments, participants were randomly assigned to see either a vignette about the implementation of a policy (A or B) or a vignette about a decision to compare them in an A/B test (randomized experiment).

Free Response Coding. We used a conventional content analysis approach (30) to inductively and iteratively develop a codebook based on initial review of participant comments from study 1. Two independent coders then applied the codebook to all responses from these four experiments (average inter-rater reliability across four studies coded, Cohen's κ = 0.83), resolving disagreements by discussion. Two new codes were added for studies 4 and 5.

Data and Materials Availability. All materials, methods, preregistrations, analysis scripts, and data are available in the *SI Appendix* or at OSF, https://osf.io/5y4f9/ (31).

ACKNOWLEDGMENTS. We thank Anh Huynh for research assistance and Sanjay Srivastava, Daniel Simons, Jean-Francois Bonnefon, and seminar participants at the Wharton School and the Institute for Advanced Study in Toulouse for comments on this work.

1. Baldassarri D, Abascal M (2017) Field experiments across the social sciences. *Annu Rev Sociol* 43:41–73.
2. Haynes L, Service O, Goldacre B, Torgerson D (2012) Test, learn, adapt: Developing public policy with randomised controlled trials (Cabinet Office Behavioural Insights Team, London). Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf. Accessed December 8, 2018.
3. Greiner DJ, Pattanayak CW (2012) Randomized evaluation in legal assistance: What difference does representation (offer and actual use) make? *Yale Law J* 121:2118–2214.
4. Institute of Medicine Roundtable on Evidence-Based Medicine (2007) *The Learning Healthcare System: Workshop Summary*, eds Olsen L, Aisner D, McGinnis JM (National Academies Press, Washington, DC).
5. Connolly P, Biggart A, Miller S, O'Hare L, Thurston A (2017) *Using Randomised Controlled Trials in Education* (Sage, London).
6. Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard Business Rev*. Available at https://hbr.org/2017/09/the-surprising-power-of-online-experiments. Accessed December 8, 2018.
7. Fisher B, et al. (2002) Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *N Engl J Med* 347:1233–1241.
8. Strauss V (April 23, 2018) Pearson conducts experiment on thousands of college students without their knowledge. *Washington Post*. Available at www.washingtonpost.com/news/answer-sheet/wp/2018/04/23/pearson-conducts-experiment-on-thousands-of-college-students-without-their-knowledge/?utm_term=.a0cd06da42d1. Accessed December 8, 2018.
9. Meyer MN (2015) Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colo Tech Law J* 13:273–331.
10. Meyer MN, et al. (2014) Misjudgements will drive social trials underground. *Nature* 511:265.
11. Goel V (June 30, 2014) Outcry greets Facebook's emotion test. *NY Times*, Section B1.
12. Tavernise S (September 8, 2015) Study of premature babies raises debate over risks and ethical consent. *NY Times*, Section A16.
13. Rosenbaum L (2016) Leaping without looking—Duty hours, autonomy, and the risks of research and practice. *N Engl J Med* 374:701–703.
14. Bazerman MH, Moore DA, Tenbrunsel AE, Wade-Benzoni KA, Blount S (1999) Explaining how preferences change across joint versus separate evaluation. *J Econ Behav Organ* 39:41–58.
15. Wennberg JE (1984) Dealing with medical practice variations: A proposal for action. *Health Aff (Millwood)* 3:6–32.
16. Faden RR, Beauchamp TL, Kass NE (2014) Informed consent, comparative effectiveness, and learning health care. *N Engl J Med* 370:766–768.
17. Gelinas L, Wertheimer A, Miller FG (2016) When and why is research without consent permissible? *Hastings Cent Rep* 46:35–43.
18. Department of Health and Human Services (2018) Federal policy for the protection of human subjects. 45 C.F.R. § 46.116(f).
19. Stiles A (2009) Literature in *Mind*: H. G. Wells and the evolution of the mad scientist. *J Hist Ideas* 70:317–339.
20. Rozenblit L, Keil F (2002) The misunderstood limits of folk science: An illusion of explanatory depth. *Cogn Sci* 26:521–562.
21. Cico SJ, Vogeley E, Doyle WJ (2011) Informed consent language and parents' willingness to enroll their children in research. *IRB* 33:6–13.
22. Clarke D, Gerrity G, Stametz R, Young A, Davis D (2017) Organizational learning in an integrated health system: Informing operations for a learning health care system. Poster presentation, Advancing Learning Health Systems through Embedded Research, 2017 Health Care Systems Research Conference, March 21–23, 2017, San Diego. Available at www.hcsrnmeeting.org/2017/local/uploads/content/files/HCSRN%202017%20Posters_2.pdf. Accessed December 8, 2018.
23. Psek W, et al. (2016) Leadership perspectives on operationalizing the learning health care system in an integrated delivery system. *EGEMS (Wash DC)* 4:1233.
24. Al-Ubaydli O, List JA, Suskind DL (2017) What can we learn from experiments? Understanding the threats to the scalability of experimental results. *Am Econ Rev* 107:282–286.
25. Stolberg M (2006) Inventing the randomized double-blind trial: The Nuremberg salt test of 1835. *J R Soc Med* 99:642–643.
26. Fisher RA (1935) *The Design of Experiments* (Oliver & Boyd, London).
27. Manzi J (2012) *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society* (Basic Books, New York).
28. Kass NE, et al. (2013) The research-treatment distinction: A problematic approach for determining which activities should have ethical oversight. *Hastings Cent Rep* 43(Suppl 1):S4–S15.
29. Asch DA, Ziolek TA, Mehta SJ (2017) Misdirections in informed consent—Impediments to health care innovation. *N Engl J Med* 377:1412–1414.
30. Hsieh HF, Shannon SE (2005) Three approaches to qualitative content analysis. *Qual Health Res* 15:1277–1288.
31. Meyer MN, et al. (2019) Data from "The A/B effect: Objecting to experiments that compare two unobjectionable policies or treatments." Open Science Framework. Available at https://osf.io/5y4f9/. Deposited April 20, 2019.