Comment

ROBERT NICKELSBERG/GETTY

Natural field experiments might involve studying how shoppers respond to in-store changes, such as moving where products are displayed.

Make science more reliable: study people as they go about their lives

John A. List

Researchers have long discussed the replication crisis. Now it's time to tackle another problem in the behavioural sciences: generalizability.

Public policy is full of initiatives that did not work out as hoped. Take Scared Straight, a programme run by more than 30 US states between 1978 and 2015 that aimed to dissuade at-risk teenagers from becoming hardened criminals by bringing them face-to-face with people incarcerated in maximum-security prisons[1,2]. The programme was extended after a pilot project, the subject of a 1978 documentary, found that 80–90% of teenage participants stayed out of trouble. But the intervention did not work when scaled up. In some places, criminal behaviour among teenagers even rose.

Similarly, many childhood-development interventions that have proved effective in one place have failed to deliver comparable results elsewhere. Deworming children in schools, for instance, substantially reduced absenteeism in Kenya but has shown mixed or weaker effects in other settings. School meal programmes in Burkina Faso increased student attendance but had limited impacts on outcomes in other countries[3].

This generalizability problem arises, in part, because human behaviour differs across populations and situations. People live in complex social environments in which labels, stakes and scrutiny shape every decision. But those contexts are often overlooked. Conventionally, research participants have come from Western, educated, industrialized, rich and democratic (WEIRD) populations[4], which are unrepresentative of other groups on measures ranging from visual perception to moral reasoning and cooperation. What might work for them might not apply to other populations.

In my view, one solution to the problem is to use a greater number of natural field experiments. In such studies, participants go about their everyday activities unaware that they are being observed by researchers, while some feature of their environment is varied. By studying people in their natural setting, assuming that strict ethical rules are followed, researchers can be more confident that their findings will be relevant to that group.

Three developments make scientists better placed to take advantage of such approaches. First, attention on the replication crisis in academia has coincided with a growing understanding that results often fail to generalize beyond the narrow populations typically recruited for laboratory studies. Second, the technology sector is running thousands of natural field experiments to obtain reliable information about their customers, establishing infrastructure and methods that academics can use. Third, a growing body of research into generalizability has provided frameworks for predicting when and why results will fail to apply across populations and settings[5,6].

Here I outline how academics can embed natural field experiments in their work.

A three-stage problem

Difficulties in replicating studies have been recognized across the sciences, from social science to biomedicine. Reforms in how research is done can help researchers to generate more-reliable results and repeat others' work more easily. This includes pre-registration of hypotheses and methods, larger samples, open data and transparent reporting. But, in fields related to human behaviour, replication only requires researchers to obtain the same result with the same kinds of people in the same setting, not to test whether it holds for people elsewhere.

Generalizability problems arise at three stages of experimental design.

Stage one: population selection. Before any experiment is conducted, researchers must select a population from which to draw study participants. A psychologist might recruit undergraduate students, whereas a medical researcher might choose from people at an academic hospital. When the population chosen for the study differs from those who will ultimately be affected by its results, the findings might not translate. Historically, clinical trials have recruited predominantly middle-aged white men, whose results were then applied to women and other groups for whom treatment effects might differ substantially[7].

Stage two: participant selection. In standard behavioural experiments, participation requires consent, and consent requires awareness. Once potential volunteers learn about a study, some agree and some decline, and this decision is rarely random[7]. Consider a lab experiment that pays volunteers US$20 to show up. Those who respond are likely to have flexible schedules, feel comfortable in academic settings and value the payment enough to participate. Such self-selection potentially skews the participant pool. In Scared Straight, this manifested in two ways. First, the volunteers were teenagers who both wanted to change and were willing to be filmed. Second, outcome data were collected through letters sent to parents, and those who had good news about their child were more likely to reply[2].

Stage three: situation selection. Every study of human behaviour involves creating an experimental setting in which to observe a participant. Situational bias arises when that context differs from how an intervention would occur in real life. This might include

GETTY

Laboratory experiments, in which people know that they are being studied, can be a good complement to natural field experiments.

differences in the level of scrutiny that volunteers are under, the magnitude of the stakes and the social cues that surround decisions.

In one study that I conducted, trading-card dealers interacted with their customers when they were knowingly under scrutiny from researchers. I then compared their behaviour with dealers working at a market who were unaware that they were being observed[8]. When dealers knew that they were being watched, they offered cards of higher quality than buyers could verify on the spot – a costly act of reciprocity unrelated to any prospect of repeat business. On the market floor, by contrast, reciprocity was strategic: generosity was extended only when reputation and repeat business made it economically rational.

As this shows, in many real-world settings, decisions occur inside webs of reputation, relationships and consequence. Generalizing from settings that mute these influences can lead to erroneous inference and flawed policymaking[9].

Observe natural behaviours

Natural field experiments can bypass many of these problems. Because people do not know that they are in an experiment, issues of participant selection (stage two) are absent. And because the experiments are conducted in the settings that researchers wish to learn about, with individuals shopping, working or donating as they normally would, the problems of situational mismatch (stage three) vanish, too[10].

Because the population is still selected by the researcher, this approach does not guarantee that stage one problems are overcome – the same result might not be found in a different community. But these experiments do clearly define the population to which the result applies. If results diverge across settings, then the reason is clear: the populations differ.

As chief economist at the retail chain Walmart, based in Bentonville, Arkansas, I know that the findings of natural field experiments that I conduct are reliable for Walmart customers, but might not apply to shoppers at Amazon, for example. By contrast, lab experiments that have participant selection can display divergent results across studies. These results could reflect genuine population differences, variations in participant selection, or both, and the researcher cannot easily disentangle them.

Natural field experiments can be used to ask a variety of questions in many sectors. A researcher studying charitable giving might vary the contents of a donor letter sent to households. A psychologist studying honesty might leave wallets in public places and measure return rates across neighbourhoods. At Walmart, my team is running natural field experiments with more than 6,000 suppliers to test which factors will most effectively incentivize suppliers to reduce their carbon emissions.

However, there are limitations. Natural field experiments cannot be applied to every research question. Some interventions, in psychotherapy, say, or classroom instruction, inherently require people's awareness. Some processes, such as private deliberations, are unobservable without asking. And there are strict ethical limits that dictate when it is and is not appropriate to use natural field experiments.

Consider ethics

Natural field experiments should not expose participants to more than minimal risk, or to experiences that they would not normally encounter. An experiment that varies the wording of a charitable appeal, the format of an energy bill or the timing of a nudge to schedule a medical appointment exposes participants to experiences in the normal range of what they would have encountered without the study. By contrast, a study that feeds negative content to users of an online platform – a manipulation carrying psychological risk beyond their normal experience – can cross the ethical boundary.

"In many real-world settings, decisions occur inside webs of reputation, relationships and consequence."

All behavioural experiments in the field should follow existing ethical frameworks governing human research. For example, *The Belmont Report* (the foundation of research ethics in the United States) notes that research involving incomplete disclosure is justified only when three criteria are met: it is needed to accomplish research goals, there are no or minimal undisclosed risks, and there is an adequate plan for debriefing when appropriate[11].

Crucially, *The Belmont Report* instructs institutional review boards (IRBs) to "distinguish cases in which disclosure would destroy or invalidate the research from cases in which disclosure would simply inconvenience the investigator". The US Common Rule codifies the principles of *The Belmont Report* into federal law, and the World Medical Association's Declaration of Helsinki (which defines ethical principles for medical research) similarly permits exceptions to informed consent[12]. As such, IRBs at universities, hospitals and research institutions routinely approve natural field experiments when the intervention falls inside the range of ordinary experience, the risks are minimal and there is no practical alternative that preserves the study's scientific validity.

Researchers must also weigh up potential harm beyond the risks for individual participants[7]. As an example, consider a 2014 natural field experiment in Montana that sent voters official-looking leaflets ranking judicial candidates on an ideological scale. No participant faced more than trivial individual risk, yet the intervention's scale could plausibly have shifted the election. Most would view this as societal harm[13].

Researchers, IRBs and partner organizations should therefore always ask whether an intervention could materially affect collective outcomes, such as elections, prices or public-health behaviour[7,14]. Smaller pilot studies, debriefing of affected communities and pre-registered stopping rules offer practical safeguards when the answer is uncertain.

Requiring awareness as a precondition for research does not necessarily protect participants. Indeed, when awareness of being studied distorts behaviour, it can yield misleading evidence that can produce bad policies, affecting many more people. The teenagers harmed by Scared Straight were not harmed by a natural field experiment, but by a programme that was scaled on the basis of ungeneralizable evidence. In such cases, the ethical question extends beyond whether running a natural field experiment affects participants, and asks whether refusing to run one inflicts greater costs on everyone the resulting policy will ever touch.

Boost natural field experiments

Laboratory and natural field experiments can be complementary. Scientists can benefit from using both approaches together.

Researchers who routinely turn to the lab should ask whether their question can also be studied in the field and design research programmes that use both to compare results. If that is infeasible, researchers designing a lab experiment should try to predict the ways in which real-world settings might differ and build as many of those features into their protocol as possible[5]. For instance, a researcher studying charitable giving in the lab might introduce time pressure that mimics a real solicitation or ensure that donation decisions are anonymous.

Often, natural field experiments require working with an organization that is willing to embed research in its operations. Doctoral programmes should prepare researchers to negotiate with organizations and understand how they operate, be it a retailer trialling a new product label, a hospital testing appointment reminders or a government agency assessing the wording of a tax letter.

Journals and funders should treat natural field experiments as a valuable way to provide evidence about when and where a research finding applies. When there are policy implications, journal editors should urge researchers to discuss how findings would be tested in natural settings. Authors should also highlight the conditions under which results might fail to generalize.

Funders should create dedicated pathways for the longer timelines and partnership costs that field experiments require. Finding and negotiating with a partner organization can take from six months to one year. The experiment might need to run for months to capture variations in behaviour across seasons or policy changes. A three-year funding cycle or a seven-year tenure clock can accommodate at most a handful of such projects, whereas a lab-based researcher might complete a dozen studies in the same period. And time is needed to work with IRBs to ensure that experiments meet high ethical standards.

Emerging tools make a shift to natural field experiments increasingly cost-effective. Artificial-intelligence models now permit researchers to test whether people in some subgroups respond to an intervention differently than do most participants, providing crucial generalizability evidence before a programme is scaled up. And digital platforms operated by technology firms allow researchers to randomize interventions across millions of users simultaneously, at negligible cost.

The replication crisis raised the important question of whether researchers can trust findings in the settings in which they were produced. Yet, another deep question that matters for policy, business and human welfare is whether the findings mean anything beyond those settings. It is time for researchers to embrace the methods that help us answer it.

The author

John A. List is a professor of economics at the University of Chicago, Chicago, Illinois, USA; at the Australian National University in Canberra, Australia; and chief economist at Walmart in Bentonville, Arkansas, USA.
e-mail: jlist@uchicago.edu

1. Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E. & Lavenberg, J. G. *Cochrane Database Syst. Rev.* **2013**, CD002796 (2013).
2. Finckenauer, J. O. *Ann. N. Y. Acad. Sci.* **347**, 213–217 (1980).
3. Vivalt, E. *J. Eur. Econ. Assoc.* **18**, 3045–3089 (2020).
4. Henrich, J., Heine, S. J. & Norenzayan, A. *Behav. Brain Sci.* **33**, 61–83 (2010).
5. List, J. A. *Nature* **626**, 491–499 (2024).
6. Al-Ubaydli, O. & List, J. A. *On the Generalizability of Experimental Results in Economics* (Handbooks of Economic Methodology, 2015).
7. List, J. A. *Experimental Economics: Theory and Practice* (Univ. Chicago Press, 2026).
8. List, J. A. *J. Polit. Econ.* **114**, 1–37 (2006).
9. Levitt, S. D. & List, J. A. *Science* **319**, 909–910 (2008).
10. Levitt, S. D. & List, J. A. *J. Econ. Perspect.* **21**, 153–174 (2007).
11. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report* (US Department of Health, Education, and Welfare, 1979).
12. World Medical Association. *JAMA* **310**, 2191–2194 (2013).
13. Willis, D. 'Professors' research project stirs political outrage in Montana.' *The New York Times* (28 October 2014).
14. Bernstein, M. S. *et al. Proc. Natl Acad. Sci. USA* **118**, e2117261118 (2021).

The author declares no competing interests.

Students restore relics as part of their course at a technical college in China.

Why science needs the humanities more than ever

Xin Fan

From AI to biotechnology, today's scientific breakthroughs raise ethical and social challenges that technical expertise alone cannot solve.

A narrative has taken hold that science and the humanities are at odds. In universities around the world, investment in cutting-edge technologies often comes at the expense of retrenchment in philosophy, history, literature and the arts. Contraction of the humanities is presented as an unavoidable cost of modernization. But that 'zero sum' logic is flawed. As science and technology race ahead, the world needs humanities research to understand the reasons and implications.