Beyond Experiments

Ed Diener
Deceased. Formerly University of Utah, University of Virginia, and the Gallup Organization

Robert Northcott
Birkbeck College, University of London

Mike Zyphur
University of Melbourne

Stephen G. West
Arizona State University

Running Heads: Beyond Experiments

Send inquiries to: Michael Zyphur, mzyphur@unimelb.edu.au

Acknowledgements
We express our gratitude to the following for their helpful comments on this paper: Jonathan Butner, Michael Eid, Aaron Likens, Louis Tay, and Timothy Wilson

Abstract

It is often claimed that only experiments can support strong causal inferences and therefore they should be privileged in the behavioral sciences. We disagree. Overvaluing experiments results in their overuse both by researchers and decision-makers, and in an underappreciation of their shortcomings. Neglecting other methods often follows. Experiments can suggest whether X causes Y in a specific experimental setting; however, they often fail to elucidate either the mechanisms responsible for an effect, or the strength of an effect in everyday natural settings. In this paper, we consider two overarching issues. First, experiments have important limitations. We highlight problems with: external, construct, statistical conclusion, and internal validity; replicability; and with conceptual issues associated with simple X-causes-Y thinking. Second, quasi-experimental and non-experimental methods are absolutely essential. As well as themselves estimating causal effects, these other methods can provide information and understanding that goes beyond that provided by experiments. A research program progresses best when experiments are not treated as privileged but instead are combined with these other methods.

Beyond Experiments

Experiments can be defined as studies in which a researcher manipulates a treatment condition (putative causal factor) and compares the dependent variable scores with those from another manipulated treatment condition, often a 'no treatment' control. Assignment of people or units to treatment conditions is random and often there are attempts to control other relevant factors. Quasi-experiments manipulate the treatment condition and contain some features of experiments, but do not include randomization. The strength of experiments is that they address some threats to internal validity (see below), leading to the claim that only experiments can rule out both known and unknown confounding factors. The weakness of experiments is that they do not address other forms of validity (statistical conclusion validity, construct validity, and external validity), and indeed can make addressing them harder. We suggest that experiments are overused and overvalued in the behavioral sciences, to the detriment of scientific progress.

This view may seem heretical because experiments are often held in the highest esteem. For example, Imbens (2010) writes that "Randomized experiments enjoy a special place in the hierarchy of evidence, namely at the very top" (p. 407). Many agree, but we do not. An overemphasis on experiments can blind researchers to their many shortcomings, and to the absolute necessity of incorporating other methods.

Some scholars have begun to question the strong emphasis on experiments, and there has been a reaction against inflexible hierarchies of evidence that promote experiments as a unique gold standard, superior to other methods always and everywhere. In medicine, Jadad and Enkin (2007), in a revision of their volume on randomized controlled trials, present a greatly enhanced discussion of the biases in experiments. In economics, Nobel prize winner Deaton (2010) writes that "experiments have no special ability to produce more credible knowledge than other

methods, and … actual experiments are frequently subject to practical problems that undermine any claim to statistical or epistemic superiority" (p. 424). In epidemiology, Hernán (2018) and Hernán and Robins (2020) argue that the results from randomized experiments cannot be given a "free pass," immune from the scrutiny given to other methods. In psychology, Cook (2018) identifies 26 assumptions that a randomized experiment must satisfy to merit gold standard status; few, if any, experiments satisfy most of them. The trend in recent philosophy of science and economics is towards a more balanced view, namely that experiments have strengths but also substantial weaknesses, both in practice and in theory (Clarke, et al. 2014; Deaton & Cartwright, 2018; Hernán, Hernández-Díaz, & Robins, 2013; Krauss, 2018; Imai, King, & Stuart, 2008; Little, 2019; Worrall, 2007; Young, 2019).

Despite this, we still find areas within psychology wedded to the old thinking. Laboratory-based experiments with questionable manipulations are used to justify grand claims about real-world phenomena. But optimal scientific progress, in psychology as elsewhere, requires the intelligent use of the full range of methods, the exact mixture varying case by case and purpose by purpose.

Non-experimental and quasi-experimental methods permit causal inference—experiments have no monopoly on that. They are also essential for discovering the mechanisms that underlie causal connections. In addition, they are usually superior for examining a large number of causes in unison and estimating their relative effects and interactions (Deaton, 2010; Heckman & Smith, 1995; Hernán & Robins, 2020). And for many phenomena in the human sciences, a complex interplay of multiple influences operates in a dynamic system. The simplifications required for experiments cannot do such phenomena justice.

To be clear: we agree that experiments have an important role to play in psychology. But their proper use is as part of a cumulative research program that combines them with other methods, including real-world observations, and that does not treat them as essential or even privileged. “Science benefits from an accumulation of results across a variety of studies. Results accumulate best when the methods used to create knowledge are varied and complementary in their strengths and weaknesses …” (Reichardt, 2019, p. 5; see also Rosenbaum, 2015).

In what follows, first we describe the limitations of experiments, including various types of validity issues, problems with replication, and conceptual issues. Next, we explain the strengths of alternative methods, arguing that alternative methods are virtually always needed. Finally, we discuss when experiments are most and least helpful, answer possible objections, and present recommendations for future research.

Limitations of Experiments

Sometimes experiments cannot be conducted for ethical or practical reasons. They can be expensive and time-consuming, volunteers can be difficult to recruit, and many important behavioral phenomena simply cannot be studied by them (Cronbach, 1957). As an example, in basic research, mild emotions can be induced in experiments, but not extreme passions such as falling in love, rage, or panic—much less their consequences. In applied research, experiments cannot study the long-term effects of child abuse. In research addressing important structural or policy issues, “participants may not want to be randomized, randomization may not be feasible or not accepted in the research context, or only atypical participants may be willing to be randomized” (West et al., 2008, p. 1359). Consider the dual difficulties of (a) convincing participants to allow themselves to be randomly assigned to a religious faith-based substance abuse treatment group or a cognitive-behavioral therapy comparison group, and (b) convincing

faith-based providers to withhold their faith-based treatment from participants assigned to the secular therapy comparison condition. Given these and other difficulties discussed below, alternatives to experiments are often required.

Types of Validity

Shadish, Cook, and Campbell (2002) defined four types of validity:

1. *External validity* has two sub-forms. In basic research, can a causal effect be generalized *across* the populations, settings, and times of interest? In applied research, can a causal effect be generalized *to* the population, settings, and times of interest? For example, basic research ideally finds causal effects that hold across contexts and research participants (e.g., males and females, cultures, and settings such as home, school, and business). In contrast, applied research often addresses whether a specific treatment program delivered in a specific context (e.g., classrooms) led to specific outcomes (e.g., reduction in cigarette smoking) in a specific population (e.g., high school students; see Evans et al., 1978).

2. *Construct validity* combines two questions. Does the manipulation affect the theoretical construct of interest or some other construct? And does the dependent measure assess the theoretical outcome of interest or something else?

3. *Statistical conclusion validity*. Does an association exist, and what is the strength of the association between two (or more) variables?

4. *Internal validity*. Can we attribute differences in the dependent variable (*Y*) to differences in the independent variable (*X*)? Does *X* cause *Y*?

Although there can be some overlap between them, it is useful to consider these types separately. We will discuss each of them in turn.

Problems of External Validity

The results of a recent experiment by Yeh et al. (2018) were unexpected: "Parachute use did not reduce death or major traumatic injury when jumping from aircraft in the first randomized evaluation of this intervention." What a remarkable finding! The authors continue: "… the trial was only able to enroll participants on small stationary aircraft on the ground, suggesting cautious extrapolation to high altitude jumps." (p. 1). The study is both silly and insightful. Obviously, people survived skydiving without a parachute only because the airplane was parked on the ground. The findings were completely dependent on this context; crucially, this context has no overlap with the context of higher-altitude jumps, normally the target of generalization for a study about parachutes. The issue here applies to many experiments in the behavioral sciences. Yet the potential dependence of results on context often goes unrecognized.

Similar issues arise when generalizing results to new populations (Yarkoni, 2020). Although psychologists have long noted the danger of generalizing from undergraduates to other populations (Sears, 1986; Henry, 2008), the practice is still widespread. The National Institutes of Health has over the past two decades increasingly emphasized the use of samples that are diverse with respect to race, ethnicity, and gender. Cultural psychologists have questioned the generalization of human research in psychology that uses WEIRD participants—Western, Educated, Industrialized, Rich, and Democratic (Henrich, Heine, & Norenzayan, 2010). Despite this, most psychological research on humans has utilized convenience samples (often undergraduates) with little concern for whether these samples represent the target population of interest. Syed (2021) argues that in social psychology an overreliance on experiments serves to restrict participant diversity.

Consider next the cautionary tale of the Women's Health Initiative (WHI), an influential randomized controlled trial on hormone replacement therapy (HRT; Cagnacci & Venier, 2019).

Previous very large observational studies suggested benefits of HRT. So, a large randomized experiment – the WHI – was run. It indicated that HRT risks outweigh their benefits, a result that received broad media attention and led to the marked decline of HRT. The WHI trial has been interpreted as showing the dangers of observational studies. However, subsequent research indicated that the results of the WHI experiment did not generalize to women of all ages, to all types and administrations of HRT, and so forth. For example, (Hernán, et al., 2008) reanalyzed a large, long-term observational study that had collected data on HRT, the Nurses Health Study. They found that when they estimated the same causal effect as in the WHI, the intention-to-treat effect, and compared the effects of HRT on women 10-20 years post-menopause, the results of the observational study and the WHI randomized trial were very similar. How so? The majority of women who volunteered for the WHI trial were many years post-menopausal (average age 63), and the conclusions of the trial turned out to be largely because of this (Lobo, 2017). For younger women, HRT actually reduced all-cause mortality, and had other benefits such as increased bone density. A meta-analysis of HRT showed that it substantially reduces all-cause mortality for women under 60 (Salpeter et al., 2004). Failure to appreciate issues of external validity in this case could have, and may continue, to cost a large number of lives.

Perhaps because experiments are so highly valued as engines for causal inference, the importance of context-dependence, population-dependence, and time-dependence is often neglected (Yarkoni, 2020). We advise researchers in their own work to assess whether: the airplane is in the air; postmenopausal women are not representing those women currently experiencing menopause; and appropriate time has elapsed between treatment and assessment. Psychologists cannot ignore the external validity issue, but must examine it for the *specific* contexts and populations to which an experimental finding is meant to generalize.

Deeper Understanding of Context. We will consider context in some depth because its role in generalization has received the least attention from psychologists. A study's context may include factors that are material (e.g., a study's environment), national or cultural (e.g., USA), institutional or rule-based (e.g., how participants should behave in an academic organization), and sometimes identity-based (e.g., participants respond based on their identity as a student). The aspects of a study's context that matter in a particular case are whatever aspects influence whether the causal effect(s) in an experiment apply in the relevant target situations (Cook, 2003).

In all sciences, a configuration of background and enabling conditions provide necessary scaffolding for causal connections to occur (Mackie, 1980). In human sciences, causal connections are often highly sensitive to changes in this scaffolding, as evidenced by failed educational and public-health interventions based on previous randomized clinical trials (Cartwright & Hardie, 2012). In basic research and applied trials studying treatment efficacy, experiments in psychology are typically designed to hold constant or 'control' contextual factors. This limits the ability to test for contextual moderators. The price is little insight yielded into which background conditions are needed for causal effects to occur. Because of the presumption of control and the insensitivity of results to context, researchers rarely offer the kind of rich situational descriptions that would be required to evaluate external validity. Researchers thereby also miss the opportunity to develop deeper scientific understanding. This problem is widespread, from medicine to psychology and beyond (Rothwell, 2005).

The problem may be put more formally. General considerations of evidence dictate that a claim of external validity requires a three-part justification (Cartwright and Hardie, 2012). We need good reason to believe: (1) that the relevant causal connection was operative in the old context, i.e., in the experiment; (2) that it will be operative in the new context, i.e., in the context

of application; and (3) that the relevant support factors will also be present in the new context. If we lack good reason to believe even one of these conditions, the justification for external validity fails. Crucially, the rigor of an experiment's design speaks only to the first of these conditions. The latter two conditions can be justified only by knowledge of the new context, which goes beyond what a given experiment and its associated write-up provide. Therefore, it is impossible for an experiment alone to justify a claim of external validity.

For this reason, it is no surprise that extrapolating the results of social and behavioral science experiments is notoriously unreliable (Levitt & List, 2007; Möllenkamp et al, 2019; Reiss, 2008). To believe that an experiment's findings are externally valid requires accepting that they are transferable or exportable to target situations. This is made more challenging by experiments' often contrived contexts and manipulations (see section on construct validity below). The problem of external validity will often be more pronounced for laboratory experiments than for field studies: recall that the agricultural experiments championed by Fisher used an open system randomization model in which such contextual factors as rainfall and hours of sunlight were *not* controlled. But external validity can be a challenge even for field studies, especially when additional controls are implemented. Insights about external validity usually come from wider knowledge, garnered from other methods and sources of expertise that are contextually sensitive (Deaton & Cartwright, 2018). Studies can be made more relevant to real-world contexts through their design as well as through description in write-ups, but experiments typically lack these.

A laboratory experiment allows causal inference without the apparent need to attend to context, which is meant to be one of its advantages. However, this feature becomes a *dis*advantage when trying to extrapolate an experiment's results to a new environment, because

without any properties of typical real-world situations the decontextualized study has no clear link to a real-world situation.

Some areas of basic research in traditional experimental psychology, which have placed minimal emphasis on external validity (Mook, 1983), have limited their scientific progress because of a lack of attention to context. A substantial body of research on verbal learning and memory was conducted using nonsense syllables (Hall, 1971); many of these findings were discarded in light of newer research in the area of cognitive psychology (Neisser, 1967; 1976), which observed the original findings did not generalize to the learning and memory of meaningful stimuli (e.g., words) that characterize human learning in everyday life. Gibson (1950) decried earlier research in perception, particularly work on perceptual illusions, in which the person (or animal) was constrained by making only highly controlled, limited information available. Gibson (1966) instead emphasized the far greater information available when the person was actively moving in or interacting with the environment, as occurs in everyday life (Carello & Turvey, 2017). McBeath et al. (1995) demonstrated the additional critical perceptual information available to baseball outfielders who are running to locate and catch fly balls. Finally, current fMRI studies permit the observation of neural activity in tightly controlled experimental conditions. However, those conditions have caused a variety of serious problems, including the erroneous inference that legitimate cortical activity is ‘noise’ by failing to appreciate brain-wide representations of behavior-environment interactions (see Stringer et al., 2019). They have also led researchers to fail to measure ongoing ‘twitches, blinks, and fidgets’ that explain a surprising amount of the variation in fMRI results (see Drew et al., 2019). Thus, due to the contrived and constrained nature of the experimental context, the applicability of the experimental results to field contexts may be undercut *by design*.

Is there a statistical short-cut to ensure external validity? Some statistical techniques track how the treatment effect within a trial varies with particular combinations of causes. The motivation is that this knowledge can shed light on external validity. In a series of papers, Pearl and Bareinboim give formal, generalized versions of this procedure (e.g., Bareinboim & Pearl 2013; Pearl & Bareinboim 2014), applying methods from causal graph theory and structural causal modelling. Their work describes what information is inferable about new populations from trial results, given certain assumptions about causation. It confirms that such inferences require extensive knowledge of probabilistic and causal dependencies in both the study's original population and in the new populations to which the results might be applied. Similarly, sophisticated econometric techniques can track interacting variables present in both original and target environments, in order to estimate outcomes in the target environment from differences in the distributions of these interacting variables between the two environments (Crump, Holtz, Imbens, & Mitnik, 2008; Muller, 2015). Effectiveness experiments in which treatments are tested in a heterogeneous or modal sample of the contexts of interest (Shadish et al., 2002), under less controlled conditions, offer a stronger basis from which to generalize. But to secure external validity, all of these techniques require extensive knowledge of the target environment (Khosrowi, 2019). There is no short-cut.

Individual and Population Differences. Individuals differ, sometimes strongly. In psychology, based on our statistical models (e.g., analysis of variance), historically we assumed that treatments had a constant effect, affecting all participants to the same degree. More recently, we have recognized that experiments attempt only to estimate average causal effects (Imbens & Rubin, 2015; West & Thoemmes, 2010), and that causal effects for individuals are distributed around that average. Such average effects cannot be particularized to specific individuals

(Molenaar, 2004; West & Thoemmes, 2010). Many or even most participants might not be affected by a treatment; some may experience effects of opposite sign to the population average. Bespoke knowledge of individual causal effects may often be more important epistemically than a statistical average causal effect. But identifying the characteristics of the individuals who are more or less responsive to a treatment can rarely be done by the experimental method alone.

Timing of Measurement. A final concern is the time-course of the effects of an intervention. Reichardt (2019; Reichardt & Gollub, 1987) emphasizes the sensitivity of observed treatment effects to the lag between treatment and measurement of the dependent variable. Experiments are expensive, which often limits their duration, meaning we cannot know whether and to what extent an intervention is useful in the long run. Consider the effects of corporal punishment. There is mixed evidence about whether physical punishment reduces anti-social behavior in the short run, with some studies indicating that it does. Yet, extensive research also indicates that in the long run anti-social behavior is consistently increased by it (Heilmann et al., 2021; Smith, 2006). Treatment effects can systematically increase or decrease over time, or even reverse direction entirely.

Participants can age, and important changes in context occur (e.g., transition from childhood to adolescence). When measurements are made can make a crucial difference. This might seem obvious, but it can influence external validity in profound ways by making findings justified only for the timeframe covered by a study's design. Unfortunately, psychologists often collect data over short time periods and then assume that the treatment effect generalizes over time, without knowing whether it actually does. Supplementing experimental results with those of observational studies that can model both short-run and long-run effects separately, addresses this issue (Zyphur et al., 2020; Shamsollahi, Zyphur, & Ozkok, in press). Given that multi-

billion-dollar interventions can hinge on long-run effects, the typical experimental design of short-run interventions and assessments is sorely lacking.

External validity is an issue facing all research designs, but it is especially acute with experiments. Experimental contexts are often unrepresentative of target contexts. Participants in an experiment are rarely sampled from a population of interest. Further, participation in an experiment may require high levels of commitment and motivation, and volunteers are often more conscientious and educated than non-volunteers. Although non-experimental designs may also be demanding of participants, many experiments require greater commitment. And the duration of an experiment may be limited, so that the temporal course of treatment effects is unknown. Quasi-experiments and observational studies conducted in situ are often *much better* for matching the context, participants, and time course with the targets of generalization.

Problems of Construct Validity

For in psychology, there are experimental methods and conceptual confusion. The existence of the experimental method makes us think that we have the means of solving the problems which trouble us; though problem and method pass one another by. – Wittgenstein (1958), p. 232

Carefully conducted experiments that satisfy the design's underlying assumptions are high in internal validity. They license a specific causal conclusion: something about the manipulation caused the observed changes in the outcome. Yet, there is a danger that experiments in themselves do not directly address. The constructs that comprise our theoretical independent variables in psychology typically have no single definitive way of being operationalized. And the constructs that comprise our theoretical dependent variables also typically have no single definitive way of being operationalized. This creates a potential gap. On one side, there is the hypothesized relationship between the theoretical constructs (the causal

process being tested), and on the other side, there is the way those constructs are operationalized by an experiment's intervention and measurement activities.

Modern experimental methods based on randomization were developed and popularized initially in agriculture (Fisher, 1935), and later in medicine. In these domains, often the key question is: does a particular concrete treatment have a specific beneficial effect? Does a fertilizer increase the crop yields of barley, for example, or does a medication reduce the probability of death in covid-19 patients? The manipulated treatment variable closely mirrors the real-world intervention that the experiment is designed to test, and the dependent variable is the specific outcome of interest. But matters are different in most psychological research. Consider, for example, whether frustration causes aggression. In an experiment, what is tested is whether making people experience frustration in a particular way (e.g., a specific insult) leads them to give aggressive responses of a specific type (e.g., delivery of electric shocks). Although the experiment might justify inferring a causal connection between these particular operationalizations, it still leaves uncertain whether there is a causal connection between the more abstract concepts of frustration and aggression—which is the true question of interest. A community of researchers must decide whether a particular operationalization really does exemplify the target theoretical construct. Contrast this to the agricultural experiment with fertilizers. Experiments as they historically developed in agriculture and medicine were *not* designed for the study of abstract constructs in the way that psychologists typically use them.

In work with abstract concepts, it is helpful to have other validated measures of the theoretical independent and theoretical dependent variables, to provide evidence that these variables are indeed assessing the target construct (convergent validity). Checks on the manipulation provide this evidence for theoretical independent variables; alternative assessments

of the construct provide this validity evidence for the dependent variables. This allows research to support a causal sequence that runs from the manipulation through the conceptual independent variable, to the conceptual dependent variable, to the operational dependent measure. Fiedler, McCaughey, and Prager (2021) offer an impassioned cry for checks on the manipulation in both experimental and quasi-experimental research. They find that in five excellent journals, even in the best studies only 50 percent included checks on the manipulation.

Researchers also need to assess possible third variables affected by the manipulation of an independent variable. For example, an insult used to manipulate the conceptual independent variable of frustration might also lead to enhanced physiological arousal, facilitating more extreme responses in general, including more extreme aggressive responses (see West, Kwok, & Liu, 2014). In their review, Fiedler et al. (2021) found that virtually no studies checked for such possible confounding. They argue that such checks are key for theory building: they can be leveraged in a mediational model that probes the relative effect of a treatment through each mediator to a dependent variable (MacKinnon, 2008; Vanderweele, 2015; West & Aiken, 1997). Fiedler et al. argue that such mediational analysis is crucial, suggesting that "no manipulation can be expected to affect only a single [theoretical] IV" (2021, p. 3). They suggest that manipulation checks should be included in all research. Without checks on the manipulation, as well as on other potential confounding variables, it is not possible to know what it is about a manipulation that produced the effect.

In reality, it is often impossible to manipulate an abstract psychological construct and change nothing else, so experiments are vulnerable to possible third-variable explanations. Experimental treatments are usually not toggle-switch activities in which the intended variable is turned on or not and nothing else is manipulated, as with fertilizers. Suppose that participants

watch a movie, for example. This can induce a mood change, but it can also induce thoughts about sociability, create arousal, or affect the dependent variable in other ways. It may be these unintended effects that are causing the observed outcome. Just as non-experiments do, experiments should present evidence that unintended effects of any manipulations are not a problem—but we do not see much of this in psychology. Because of their perceived strength, experiments are often given a 'free pass', with little scrutiny for potential confounders.

Chester and Lasko (2020) provide an informative review of manipulations in social psychology. Examining 348 experimental manipulations in the *Journal of Personality and Social Psychology*, they found that very few were checked for validity. Most were created 'on-the-fly' without a history of use and without validation studies such as pilot testing. Chester and Lasko recommend that experimenters use validated and standardized manipulation protocols, study the effects and duration of the manipulation, and estimate the manipulation's effects on multiple possible constructs within the target theory. They found that these recommendations are rarely if ever comprehensively followed.

Manipulation checks in mood experiments cast doubt on whether the intended moods were actually induced. For example, Diener, Oishi, and Cha (2021; see also Joseph et al., 2020) found that people in negative mood experimental conditions were actually in positive moods after the induction, albeit less positive than before the manipulation and more negative than in positive mood experimental conditions, but nonetheless still in a positive mood. Thus, what is learned from so-called negative mood inductions is often actually about milder positive moods. The construct validity is dubious.

Finally, experiments may also induce suspicion of experimenters' intent, or induce other motivations such as attempting to provide the experimenter with the desired results (for a review

see Kruglanski, 1975). Experiments are arguably especially vulnerable to this problem. They require high levels of motivation and co-operation from participants, and they often intervene in participants' lives more substantially than other methods do.

In sum, because of the gap between experimental operations and the concepts they are meant to represent, imputing causality to a particular concept is usually less valid than assumed. Yet, some researchers do not spot the problem. They believe their own conceptual labels, and do not look at the many other factors that their experimental treatment may also be affecting.

Problems of Statistical Conclusion Validity and Replication

Shadish et al. (2002, p. 45) detail the many ways in which the detection of causal associations can be impeded. One of the most important is low statistical power, in other words a low probability that a study will detect an effect of a specified size even though the effect, in fact, exists. Rossi (2013) documented the low statistical power of studies in leading psychology journals; only recently has a priori calculation of statistical power been emphasized as a standard in psychology (Appelbaum et al., 2018). It is often difficult for experiments to achieve the sample sizes necessary for satisfactory levels of statistical power, because it can be difficult to recruit participants who will agree to all aspects of the treatment, will participate in the (perhaps inconvenient laboratory or clinic) controlled setting established by the experimenter, and will commit to completing all of the assessments. In contrast, quasi-experimental and non-experimental designs can sometimes achieve much larger samples. For example, children in school systems throughout a state may be assigned (or not) to a school-based lunch program based on family income, and their (anonymized) standardized tests of achievement and school grades used as the dependent variables.

The so-called replication crisis in psychology also highlights problems of statistical conclusion validity that impact experiments in particular: for experiments may be less replicable than non-experiments. Nosek and colleagues (Open Science Collaboration, 2015) attempted close replications of 100 published psychological studies. In many cases, they used the original study materials and consulted with the original authors to get the procedures correct. The studies were from across psychology, published in three highly prestigious journals. But in only 39% of the cases was the major finding of the original study replicated—far less than half.

We classified each of the studies from the Open Science project as experimental or non-experimental, so as to investigate replicability as a function of study design. We found that only 27% of the 66 experimental studies replicated, whereas 71% of the 17 non-experimental studies did. These results preliminarily suggest that non-experimental findings are more replicable. We invite others to repeat this exercise more formally, both with the Open Science studies and with other studies of replication of psychological research.

For some of Nosek et al's studies, the original authors expressed concern that the replication attempt lacked statistical power or that it did not closely reproduce the original study's protocol. Ebersole and colleagues (2020) re-replicated these studies in particular. They instituted stringent standards for their study protocol, such as obtaining feedback on the methods from the original authors, and having reviewers approve the replication methods. They conducted each re-replication in an average of 6.5 laboratories, and used large sample sizes to ensure statistical power. Nine of the studies were experiments, and six of these yielded relatively clear findings. Of these six, five failed to replicate, thus speaking strongly against alternative explanations for the non-replications by Nosek and others.

These are striking and alarming results. In articles chosen from some of the most selective journals in psychology, and therefore presumably judged to be rigorous by some of the field's top reviewers and editors, barely more than 25% of experimental studies replicated! Non-experiments did much better. Although the selection of studies by the Open Science project was not random, the roughly 17% success rate of very close replications of experiments by Ebersole and colleagues (2020) raises serious concerns. If our goal is to be able to generalize the results of experiments to other populations and settings, then getting similar results in close replications should be a low bar. The onus is on researchers to demonstrate that their experiments are replicable before their findings can be taken seriously.

One potential solution is greater use of conceptual replications after an effect has been established (presumably by preliminary close replications), with new participant populations, new manipulations of the independent variable, new measures of the dependent variable, and new settings. These conceptual replications should limit the number of features of the original experiment that are altered, facilitating interpretable results. Such conceptual replications can check on context effects and generalizability. Unfortunately, they are rare.

Shadish et al. (2002) have argued for the primacy of literature reviews of multiple studies using multiple methods. Supporting this, recent work in bioinformatics and epidemiology shows that replicability is increased by combining many 'messy' or 'dirty' datasets for meta-analysis (Cahan & Khatri, 2020; Haynes, Tomczak, & Khatri, 2018; Vallania, et al., 2018; Sweeney, et al. 2017). A plausible message here is that reliable causal and other signals that facilitate external validity are more likely to come from noisy data that closely reflect the natural world, rather than from controlled experiments.

What is not mentioned in many current discussions is that failures to replicate may simply reflect the nature of the world. Psychological phenomena may just be very sensitive to small differences in the timing of measurements, specifics of a manipulation, participant selection, or other contextual factors (McShane et al., 2019). If so, problems of external validity will inevitably be endemic (Northcott, in press; Stroebe and Strack, 2014). They will be worse in experiments, if attempting to minimize noise through experimental controls makes it more likely that observed effects do not replicate. Issues of statistical inference also arise given that different criteria are used to infer replication, and given that effect sizes are rarely known a priori, leading to underpowered studies (Anderson, Kelley, & Maxwell, 2017; Anderson & Maxwell, 2016). Steiner et al. (2019) present a causal analysis, and Brandt et al. (2014) present a substantive analysis, of features that may differ between original and replication studies. These differences influence the judgments of peer researchers who adjudicate on whether something counts as a replication or not (Brandt et al., 2014).

Even if high levels of external validity are not always achievable, they are still desirable. Similarly, although quantitative average treatment effects are likely to vary by context—no matter what methods are used (McShane et al., 2019)—it is still desirable to pursue external validity from the perspective of the direction of an effect (Shadish, Cook, & Campbell, 2002; West & Thoemmes, 2010). Researchers should use whatever methods offer the best chance of estimating not only the mean but also the heterogeneity of treatment effects. That means moving beyond mere controlled experiments. In particular, external validity is improved by gathering and capitalizing on rich contextual knowledge, and this requires observational designs that by definition can, and likely will, involve ‘messy’ or ‘dirty’ datasets (McShane et al., 2019).

The non-replication of experiments has profound implications for the human sciences. First, no experimental should be considered of wide interest until it has been replicated, preferably across different laboratories, manipulations, measures, populations, and contexts. As is true for observational studies (Zyphur et al., 2020), experiments should be considered interesting leads, not definitive confirmation of causal effects (Deaton & Cartwright, 2018). Second, if other methods better enable a study or research program to attain a researcher's validity priorities in a given context, and can produce more replicable results, then appeals for funding or publication—as well as reviewers or editors who might recommend experiments—should be required to justify why experimental evidence is particularly useful.

Over the years we have periodically witnessed suggestions by reviewers that a contrived laboratory experiment might be more informative than observational studies. Could the findings of these studies, some of which have had millions of participants in real-world contexts across many years, really be better confirmed or disconfirmed by subjecting 30 undergraduate psychology students to a simulated manipulation of phenomena that had already been studied in their actual form in the real world? Experiments should not automatically be given preference or made a requirement for publication.

Problems of Internal Validity

The randomized experiment is in principle the strongest design for internal validity – but only if several assumptions are satisfied (Imbens & Rubin, 2015; Shadish, Cook, & Campbell, 2002). The first is that randomization has been properly carried out. A number of so-called experiments claim to use randomization, but actually use other methods of assigning participants to treatment conditions (Shadish, 2002). Second, participants must receive the full treatment to which they were assigned (i.e., full treatment adherence), and cannot switch to another condition

or find the treatment outside of the experiment (Sagarin et al., 2014). Third, there cannot be missing data on the dependent variable[1]. Fourth, there cannot be unmodeled variation in the treatment (e.g., variations in treatment implementation across sites, experimenters, or over time). Fifth, participants' (potential) outcomes cannot be affected by the treatment another participant has received (i.e., the stable unit treatment value assumption, Imbens & Rubin, 2015). For example, the outcomes of participants in a therapy group should not be affected by an aggressive group member, and the outcomes of participants in a laboratory experiment cannot be influenced by disclosures from a previous participant (e.g., a dormmate). Sixth, randomization is assumed to have an effect only indirectly through the active mechanism of treatment and not through any other effect (e.g., placebo effect; Hawthorne effect; see Angrist & Pischke, 2009). And seventh, the confounders that treatment and control groups experience after randomization may differ systematically. Cook and Campbell (1979) detail some of these validity issues such as experimenter bias, resentful demoralization, and compensatory equalization of treatments that may arise from participant or intervener communication; changes in the community context (e.g., introduction of an attractive new treatment program; local increases in cigarette taxes) can confound randomized experiments in clinical, educational, and health psychology interventions if treatment and control participants are differentially exposed. Being mindful of these other sources of confounders and controlling for them requires subject- and context-specific knowledge; it cannot be done by typical experimental design alone. Even masking the treatment

[1] Missing data biases estimates of causal effects when missingness is related to both the dependent variable and the treatment condition and/or moderator variable. Although bias can be reduced using modern missing data approaches, these corrections typically assume that all variables related to missingness are included in the missing data correction and that there are linear relationships with the outcome variable. When missingness is related to the potential value on the dependent variable (missing not at random), correction is unlikely to reduce bias.

condition fully from participants and interveners (often not achievable in psychological research) can help minimize only some, but not all of these problems.

If any of these assumptions are violated, the experiment will not produce an unbiased estimate of the causal effect. Indeed, under some violated conditions, the direction of an estimated average causal effect can even be reversed! These problems are not marginal or rare. They arise in the ten most cited randomized clinical trials in the world (Krauss 2018), and are also frequent in prestigious economic experiments (Young, 2019). Remedies exist, but they often involve additional assumptions that are hard to satisfy—if they are recognized at all.

Complexity and Non-Linear Dynamic Systems

Psychologists are aware of how 'the questions shape the answers' with surveys and similar research tools (Schwarz, 1999). What is less recognized is how larger categories of research tools also 'shape the answers' (Hacking, 2002). In particular, experiments commit us to a simplified conception of what exists in the world, i.e., to a simplified ontology. By sticking only to experiments, we limit ourselves.

As an analogy for the relationship in philosophy of science between method and ontology, consider a camera with a set of lenses (Morgan, 2006). One lens is transparent but the others are yellow, cyan, and magenta—the subtractive colors. Apply any one of these lenses, or different combinations of them, and the world appears to be fundamentally different. Just as different lenses lead to different images of the world, so too do different scientific methods, enabling and constraining scientists both materially and conceptually (Knorr-Cetina, 2009).

Experiments are one such method (Hacking, 1992a, 1992b). If one asks why they are an optimal research method, the answer given is that only experiments can license causal inferences by eliminating potential confounders. This answer is rather tautological because this way of

reasoning about causality and threats to inference is how experiments are defined. The price of it is commitment to an ontology in which the world is simplified into discrete groups (e.g., treatment versus control), causal effects are defined by average group differences, and no serious attention is given to the dynamic evolution of everyday complex systems.

Experiments are best suited to assessing single factors with simple and separable effects. They are based on the implicit assumption that a treatment simply has an effect or not, and the purpose of an experiment is just to find out which. They are far less suited to multi-factor problems that require tracking many complex interactions, feedback loops, and highly correlated processes and outcomes simultaneously. Yet many problems of great importance are like this, such as increasing life expectancy, educating children, improving social policy, and improving public health. What gets lost is the richness of real-world dynamics of biological, emotional, social, and political systems, which evolve in ways that may be non-linear and even chaotic, with causal effects that change over time (Thelen & Smith, 2007). These dynamic systems cannot be expected to be influenced by an intervention in a simple yes/no fashion. Instead, they typically produce unexpected effects and require ongoing longitudinal, observational analysis. Addressing this type of dynamic process requires a fundamentally different ontological orientation, wherein non-linear dynamic systems are the expectation rather than the exception. For example, Voelkle et al. (2018) contrast the very different answers provided by the results of a randomized clinical trial and a dynamic system analysis in the treatment of anxiety.

To illustrate, consider an example of non-experimental causal inference. Recent work in *Science*, *Nature*, and elsewhere examines how to distinguish correlation from causation in non-linear dynamic systems by using ‘convergent cross mapping’ (Clark et al., 2015; Sugihara et al., 2012). This method is derived from non-linear dynamic systems theory, which has shown that

reconstructing the behavior of a complex system and inferring causal effects requires only lags of longitudinal variables, because the behavior of the entire system is contained in the historical record of a subset of observed variables (Deyle & Sugihara, 2011). One key implication is that standard experiments go astray. They assume that the role of variables in a dynamic system can be observed by decomposition into independent components. But such a separability assumption is problematic because, when system components are deterministically linked, attempting to 'control' for relevant factors eliminates dynamics that are crucial for understanding the system as a whole (Sugihara et al., 2012). The result is simplistic ways of conceptualizing and studying phenomena that are followed only for the sake of applying experimental methods, wherein causality is inferred from average group differences. Known causal mechanisms may be either ignored or grossly simplified. Instead, researchers should embrace complexity and use methods that capture how a dynamic system actually functions.

Unlike traditional experiments, non-linear dynamic systems methods can reveal causal effects when observed variables are uncorrelated over time and appear to be entirely random. These methods can even be used to recover the causal effects produced in experiments (Ye et al., 2015). The importance of non-linear dynamics for understanding real-world phenomena is regularly overlooked by proponents of experiments, but these more sophisticated methods have revealed important effects (or lack thereof) in cases where experiments are impossible. Examples include: causal feedback between greenhouse gases and global temperatures in Earth's history, with dire long-term predictions for a warming planet (Van Nes et al., 2015), while ruling out the (rather absurd) possible confounder of 'galactic cosmic rays' (Tsonis, Hernandez, Basu, & Sugihara, 2015); the effects of the environment on biodiversity, ecological stability, and relationships between different species (Chang, et al., 2020; Ushio, et al., 2018; Ye et al., 2015);

the non-linear effects of temperature and humidity on influenza infections (Deyle, et al., 2016); and the effect of temperature on crime (Li et al., 2021). Based on this work, interventions can be planned that account for non-linear dynamics and changing causal effects (Deyle, et al. 2016).

Experimental methods are useful in their traditional domains of application, but they are often extended to dynamic systems for which they are less well suited. A variety of important phenomena simply do not fit the typical experimental ontology of average group differences and time-invariant causal effects. In such cases, the experimentalist can either ignore non-linear dynamics and the complex systems that produce them, or else try to shoehorn a complex non-linear world into what are typically simplistic linear experiments. A key purpose of the experimental method is to constrain and eliminate complexity, as if it were a confound rather than a fundamental and important property of natural systems.

Dynamic systems approaches are still being developed and bring with them their own assumptions, but these assumptions are milder than those of most experiments, and some of the dynamic systems approaches now have user-friendly implementations in Stata and R (Li et al., 2021; Ye et al., 2020). By treating experiments as the gold standard, proponents lose the motivation to seek out these alternatives.

The Need for Other Methods

McGrath (1981; see also Shadish et al., 2002) argued that any methodological approach brings with it an inherent trade-off: the better it attains one desideratum, the worse it may attain others. For a research program to maximize its cumulative contribution, it must vary its methods (Cook, 1985; Reichardt, 2019).

Fuller Understanding

Non-experimental methods are desirable because they help avoid many of the above validity problems, practical limitations, and conceptual oversimplifications. They also offer key positive advantages. In our introduction, we quoted Imbens' (2010) claim that experiments are at the top of an evidence hierarchy. But, depending on the goals of the research, the same might be said for ethnographic research, time-sampling studies, large-scale surveys, longitudinal data, machine learning applied to social media, or other methods. Each offers insights that the others do not.

The histories of many sciences point to the importance of non-experimental methods (see Daston, 2000; Daston & Galison, 2007). Galileo's simple observation and description of the Jovian moons revolutionized our conception of the universe. Darwin's far-ranging observations of life led to the theory of evolution largely without input from experiments. Many famous medical breakthroughs were not made with randomized controlled trials, and in some cases they could not have been: most surgical procedures, antibiotics, aspirin, anesthesia, immobilizing broken bones, and discovering that smoking causes cancer. The major public health approaches that have reduced smoking, such as cigarette taxes, were all evaluated using quasi-experimental or non-experimental methods. Chomsky's observations about language and Goodall's description of chimpanzee behavior were not based on experiments, and neither were other key findings in the behavioral sciences. Astronomy, climatology, geology, and anthropology were built without experiments. The successes of these disciplines indicate that non-experimental methods are not second-class citizens.

Detailing the strengths of these other methods is beyond the scope of our paper, but for illustration we briefly mention some advantages of longitudinal observational research. Long-term longitudinal studies of personality and health may span decades; short-term longitudinal

studies of respiration, heart rate, and motor behavior may span only minutes or hours. Longitudinal studies using experience sampling methods can indicate how variables move together within a person, and what changes precede other changes. These studies offer insights into natural behavior that illuminate causal sequences and processes. Researchers can observe phenomena as they unfold across stages of life. It is hard to imagine how experiments could replace this knowledge, yet in some parts of the behavioral sciences longitudinal studies are virtually never used.

Mechanisms and Structures

We do not only want to know what causes what; we also want to know why. Mechanisms and causal structures speak to the "why." We must understand the mechanisms at work within a system and how they relate to one another (Clarke et al., 2014). It is knowledge of mechanisms that guides why, and thus when, a causal connection holds. Without this, we can be lost, not knowing when an intervention will be effective.

Of course, mechanisms may sometimes be tested by manipulating mediators in controlled experimental settings (Spencer et al., 2005), but such 'double randomization' is done very rarely (MacKinnon et al., 2007), and furthermore the technique is likely to be inapplicable except in special cases. More generally, it is true that experiments do not as a matter of necessity inhibit the study of process. However, in many experiments, researchers theorize a process but then use a manipulation and measure an outcome without directly observing the process. In part, the issue here is that a process by definition is something that happens over time, involving a dynamic relationship among (changing levels of) predictors and outcomes. But most experiments are not designed to evaluate such dynamics. Instead, participants are randomly assigned to conditions and a few 'snapshots' of relevant variables are observed over what is typically a very limited

timeframe. Again, while this is not necessarily the case for experiments, it is the norm—perhaps because experimental designs lull researchers into a false sense of the validity of their research design and resulting inferences. To study process, the dynamical properties of predictors and outcomes must be evaluated together, within whatever timeframe is relevant for a given application of the research findings. Overall, therefore, although mechanisms *can* be explored by experiments, in practice they often either are not directly considered or are explored over a very limited time period. Other methods are likely to be more (if not much more) effective and efficient in this regard (e.g., Li et al., 2021).

While the effects of a known cause can be probed with experiments, the search for an unknown cause reverses the process and searches backwards from observed effects (Holland & Rubin, 1988; Pearl, 2009). In such cases, methods other than experiments come into play. Discovering mechanisms in dynamic systems requires the investigative strategies of decomposition and localization (Bechtel & Richardson, 2010). Decomposition attempts to explain a system's behavior functionally, in terms of interactions between independent sub-units. Localization then attempts to physically locate these sub-units. For example, to understand language processing we might distinguish between semantic and syntactic processing, and between speech comprehension and speech production, and then attempt to identify which parts of the brain are responsible for these functions. The approach is heuristic and iterative. Often, as in the language example, even though initial guesses are too simple, they serve to direct more fruitful follow-up work. The approach may also be extended to more integrated systems, in which sub-units are less functionally independent. All of this requires scientists to 'drill down'. The goal is not to establish an average treatment effect, but rather to tease out a mechanism's

structure: its parts, locations, and how they interact. Simple randomized trials are ill-suited to this task.

There are other important causal structures besides mechanisms. One is the *pathway*, when it useful to see a target system as analogous to our ordinary conception of roadways, highways, and city streets (Ross, 2021). Unlike with mechanisms, the emphasis is on some entity's route and flow. An example is anatomic pathways that capture physical routes, which in turn outline causal routes such as lymphatic pathways, blood vessels, and nerve tracts. Another example is metabolic pathways that capture a sequence of steps in the conversion of some initial metabolic substrate into a final downstream product. One common pathway model in developmental psychology is the *cascade*, wherein one step triggers the next step in a sequential fashion, constrained and stimulated in such a way as to achieve a certain overall effect or function. To discover these structures, we need deep programmatic research that includes many methods, not simply a never-ending series of experimental studies.

Inferring Causality with Non-experimental Methods

In psychology, there is often a taboo against the view that non-experimental results inform us about causality (Grosz, Rohrer, & Thoemmes, 2020. But they do. When longitudinal research, for example, repeatedly supports directional associations, and known third-variable explanations are controlled for, the causal implications should be taken seriously.

Across the sciences, many different methods besides experiments are used for causal inference: causal inference from observational statistics; qualitative methods such as interviews and ethnographic observation; questionnaires; small-*N* causal inference such as qualitative comparative analysis; causal process tracing; machine learning from big data; natural and quasi-experiments; as well as the methods applied to non-linear dynamic systems already mentioned.

Just like experiments, each of these methods has its own strengths and weaknesses, each has a community of practitioners, and each has a developed and rigorous methodological literature. It would be insular, to say the least, to reject all of them out of hand.

Pearl (2009; Pearl, Glymour, & Jewell, 2016) has developed a powerful graph-theoretic approach, which, given certain assumptions, offers a formal calculus for causal inference from statistics. Other research methods that were developed originally in other disciplines have also been, or are now being, incorporated into psychology. These include instrumental variable methods (Angrist & Pischke, 2009; Maydeu-Olivares, Shi, & Fairchild, 2020), difference-in-difference designs (Wing et al., 2018), and cross-lagged approaches that rely on various types of predictive or 'Granger' causality (see Zyphur et al., 2020). Even for the narrow purpose of estimating the effects of interventions, observational studies may perform as well or better than experimental studies (Benson & Hartz, 2000; Concato, Shaw, & Horwitz, 2000). In an extensive series of studies, Cook and his associates (e.g., Cook et al., 2020; St. Clair et al., 2016) compared the results of quasi-experiments and randomized experiments using an identical intervention, and showed that the estimates of treatment effects do not differ if similar quality standards are met (Thoemmes, West, & Hill, 2015). In some of these studies, participants were even randomly assigned to quasi-experimental versus randomized experimental designs. The prejudice that observational studies can shed no light on causality is outmoded.

Uses of Experiments

This leads us to the positive side of our story. Experiments should be profitably combined with other methods, as part of a cumulative program. For example, Bartels, Hastie, and Urminsky (2018) discuss the subtleties of integrating laboratory and field research with reference to one area, namely judgment and decision-making. Some laboratory findings turn out to have

substantial external validity, whereas others do not. A range of factors interact to determine optimal methodology: the details of different study designs; how wide a range of external validity is desired; the nature of the trade-offs between internal, statistical conclusion, construct, and external validity; and how much it is desirable or possible to integrate a study with wider findings and theory. The exact role of experiments varies accordingly, case by case.

Experiments can be useful even in the absence of external validity (Mook, 1983). They can confirm or disconfirm, in their specific context and population, the claims of a wider theory. When successful, experiments give a proof of concept: they show that there is at least one context, population, treatment, and dependent measure for which a claimed causal relation holds.

The contributions by experiments to the wider scientific project are visible only from a broader, theory-integrationist perspective (Deaton and Cartwright, 2018). In return, wider inputs can help with the much narrower goals that typically define experiments. For example, observational studies are often essential for assessing long-term effects and side-effects, and also for suggesting promising hypotheses to be tested by experiments in the first place.

For the simple case of confirming the effectiveness of an intervention, an experiment alone may be enough – assuming that the context and sample are representative enough of the target of generalization. But we hasten to emphasize that such simple cases are rare. More commonly, experiments provide their maximum value only when combined with other methods.

Objections

One objection to our arguments is: “nothing new here.” It might be asserted that psychologists already know the limitations of experiments and the importance of other methods, and already admit there is often a leap of faith when applying laboratory findings to the field. But even if psychologists know these things, many do not follow them in practice (for insights see

Cesario, 2021; Rozin, 2001). It can easily go unnoticed that a research program contains only experiments, with no other methods. A field can build a huge literature based solely on experiments about one causal factor, without exploring much how this factor interacts with others. Of course, some fields are better than others, and many researchers do use a variety of methods. So, our critique is not equally applicable to all areas of psychology or to all psychologists. Nonetheless, in our view, many times the value of non-experimental methods is acknowledged only in lip service, while studies, grants, and publications tell a different story in practice.

Psychology students learn the mantra that correlation does not prove causation. They do not learn the mantra that all methods are based on assumptions and that no method can lead to firm conclusions unless those assumptions are satisfied. A critique can often be heard after colloquia that a speaker "only presented observational and correlational data." But, although often applicable, one rarely hears the critique "they only presented experimental data."

Some readers might still be skeptical, insisting that psychologists know what we are advocating, and already do it. We invite readers to examine the situation and literature in their own fields. If experiments are not overvalued, that should be visible in readers' own research (see Cook et al., 2020), as well as in evidence that their fields are using a balanced variety of methods. But consider, for example, social psychology: experiments have somewhat of a stranglehold, even though many of them are contrived and decontextualized.

Another objection is that it is not just experiments that have limitations; other methods do as well. Of course they do. But because of the widespread presumption that experiments are valid, researchers are more optimistic and less critical about them than they should be. Experiments should not be given a 'pass.' Certain shortcomings are more common with them,

especially failures of external validity and construct validity, and ill-suitedness to investigating complex dynamic systems. Furthermore, our argument is not that other methods are always superior to experiments, only that they are required in addition to them. Every method has its strengths and weaknesses; different methods can be complementary.

A last objection is that division of labor in science is good: some researchers should specialize in experiments, and others in alternative methods. In principle, perhaps. The concern, however, is whether experimentalists will uphold their end of the bargain: how will a program started by an experimenter react to complementary and possibly contradictory results from quasi-experimental and non-experimental studies? The huge number of published experiments greatly outstrips attempts to replicate them, or to understand the mechanisms behind them, or to study their target phenomena in complex natural settings. Specialization can be helpful, but only if there are ways to ensure an overall balance. If most or virtually all researchers in an area specialize in experiments, then something has gone wrong.

Moving Forward

To demonstrate and understand a causal connection is not a discrete yes-no event. It is instead a process of accumulating various types of evidence that complement one another. To this end, we have nine recommendations:

1. Wording matters. In most instances, researchers should not state without qualification that "X causes Y" or "X is *the* cause of Y" based on the results of an experiment. This wording suggests more than is, in most cases, justified by the limited evidence. First, causal conclusions are based on satisfying each of the assumptions of the design; experimenters rarely probe the extent to which this was done. Second, there are always many causes of Y, not one. Third, experiments usually show only that X *can* cause Y, not

that in natural settings it is a major cause or even a cause at all. Finally, in most experiments the treatment produces only an average causal effect; the individual causal effect for a particular participant may be positive, zero, or even negative. Descriptions of experimental outcomes should be worded carefully with the qualifications clearly stated.

2. All methods are based on assumptions. For the conclusions of a research project to be accepted, those assumptions must be identified, clarified, and probed to the extent possible. When assumptions cannot be tested explicitly, that must be acknowledged, and ideally sensitivity analyses should be conducted
3. Research programs in the human sciences must use multiple methods, not just in principle but in practice. Usually, it will not be optimal if most studies are experiments. A variety of methods is essential, and reviewers and editors should be more open to these and less prone to asking authors to conduct an experiment.
4. Researchers must see experiments as only one method of causal inference among many. Diverse methods have been developed for inferring causality from non-experimental and quasi-experimental studies, and researchers should become familiar with them (e.g., Hernán & Robins, 2020; Li et al., 2021; Pearl et al., 2016; Sugihara et al., 2012).
5. External validity and construct validity should be considered from the start. An experiment alone provides at best weak evidence for these forms of validity; further evidence is required. Replication is an absolutely essential step, but only one step. The relevance of an experiment to real-world settings cannot be accepted at face value without corroborating research.
6. Experimental manipulations need to be validated to establish construct validity **(**Fiedler, McCaughey, & Prager, 2021). Do our manipulations in fact manipulate the theoretical

construct of interest? Do they not affect other constructs that may be potential confounders? Measures of the target construct and other constructs need to be entered into a statistical mediational analysis to probe the role of each intervening construct in producing the obtained outcome.

7. Where possible, researchers should conduct conceptual replications in which the putative theoretical independent variable is manipulated in several different ways and the theoretical dependent variable is measured in several different ways. Such conceptual replications reduce the chance that a finding is due to an unintended third-variable effect.
8. To discover underlying mechanisms and structures, usually non-experimental methods will be helpful and superior to experiments.

If funding and prestige are directed primarily to areas in which experiments can be conducted easily, the inevitable result will be a biased agenda, unhealthily distorting what kind of science is done (Grossman & Mackenzie, 2005). This can be pernicious. For example, it might be difficult to set up an experiment to test a nutrition program. This program might therefore be denied funding, even if potentially it could have more impact on public health than many randomized drug trials combined. To avoid such situations, we are calling for a comprehensive and diverse approach to research in the human sciences, one that is patient and that does not seek answers from experiments alone. Incentives for publications and grant awards should support the necessary variety of methods. This is the best way forward—for experimenters and non-experimenters alike.

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*(11), 1547-1562. https://doi.org/10.1177/0956797617723724

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1-21. https://doi.org/10.1037/met0000051

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3-25. https://doi.org/10.1037/amp0000191

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal inference, 1*(1), 107-134. https://doi.org/10.1515/jci-2012-0004

Bartels, D. M., Hastie, R., & Urminsky, O. (2018). Connecting laboratory and field research in judgment and decision making: Causality and the breadth of external validity. *Journal of Applied Research in Memory and Cognition, 7*(1), 11-15. https://doi.org/ 10.1016/j.jarmac.2018.01.001

Bechtel, W., & Richardson, R. (2010). *Discovering complexity*. MIT Press.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine, 342*, 1878-1886. https://doi.org/10.1056/NEJM200006223422506

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224. https://doi.org/10.1016/j.jesp.2013.10.005

Cagnacci, A., & Venier, M. (2019). The controversial history of hormone replacement therapy. *Medicina, 55*(9), 602. https://doi.org/10.3390/medicina55090602

Cahan, E. M., & Khatri, P. (2020). Data heterogeneity: The enzyme to catalyze translational bioinformatics? *Journal of Medical Internet Research, 22*(8), e18044. https://doi.org/10.2196/18044

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton & Mifflin Company.

Carello, C., & Turvey, M. T. (2017). Useful dimensions of haptic perception: 50 years after The Senses Considered as Perceptual Systems. *Ecological Psychology, 29*(2), 95-121. https://doi.org/10.1080/10407413.2017.1297188

Cartwright, N. (2009). Evidence-based policy: What's to be done about relevance? *Philosophical Studies, 143*, 127-136. https://doi.org/10.1007/s11098-008-9311-4

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better.* Oxford University Press.

Cesario, J. (2021). What can experimental studies of bias tell us about real-world group disparities? *The Behavioral and Brain Sciences*, 1-82. https://doi.org/10.1017/s0140525x21000017

Chang, C. W., Ye, H., Miki, T., Deyle, E. R., Souissi, S., Anneville, O., ... & Sugihara, G. (2020). Long-term warming destabilizes aquatic ecosystems through weakening biodiversity-mediated causal networks. *Global Change Biology*, *26*(11), 6413-6423. https://doi.org/10.1111/gcb.15323

Chester, D. S., & Lasko, E. N. (2020). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science.* https://doi.org/10.1177/1745691620950684

Cronbach, L. J (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684. https://doi.org/10.1037/h0043943

Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., & Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, *96*(5), 1174-1181. https://doi.org/10.1890/14-1479.1

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339-360. https://doi.org/10.1007/s11245-013-9220-9

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*, 1887-1892. https://doi.org/10.1056/NEJM200006223422507

Cook, T. D. (1985). Post-positivist *critical* multiplism. In R. L. Shetland, & M. M. Mark (Eds.), *Social Science and Social Policy* (pp. 21-62). Sage.

Cook, T. D. (2003). The case for studying multiple contexts simultaneously. *Addiction*, *98*, 151-155. https://doi.org/10.1046/j.1360-0443.98.s1.11.x

Cook, T. D, Zhu, N., Klein, A., Starkey, P., & Thomas, J. (2020). How much bias results if a quasi-experimental design combines local comparison groups, a pretest outcome measure and other covariates?: A within study comparison of preschool effects. *Psychological Methods*, *25*(6), 726-746. https://doi.org/10.1037/met0000260

Crump, R., Hotz, V., Imbens, G., & Mitnik, O. (2008). Tests for treatment effect heterogeneity. *The Review of Economics and Statistics, 90*(3), 389-405. https://doi.org/10.1162/rest.90.3.389

Daston, L. (Ed.). (2000). *Biographies of scientific objects*. University of Chicago Press.

Daston, L., & Galison, P. (2007). *Objectivity*. Princeton University Press.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*(2), 424-55. https://doi.org/10.1257/jel.48.2.424

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine, 210*, 2-21. https://doi.org/10.1016/j.socscimed.2017.12.005

Deyle, E. R., Maher, M. C., Hernandez, R. D., Basu, S., & Sugihara, G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences*, *113*(46), 13081-13086. https://doi.org/10.1073/pnas.1607747113

Deyle, E. R., May, R. M., Munch, S. B., & Sugihara, G. (2016). Tracking and forecasting ecosystem interactions in real time. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1822), 20152258. https://doi.org/10.1098/rspb.2015.2258

Deyle, E. R., & Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS One*, *6*(3), e18295. https://doi.org/10.1371/journal.pone.0018295

Diener, E., Oishi, S., & Cha, Y. (2020). *Reinterpreting mood induction experiments.* Manuscript in preparation, University of Virginia.

Drew, P. J., Winder, A. T., & Zhang, Q. (2019). Twitches, blinks, and fidgets: Important generators of ongoing neural activity. *The Neuroscientist, 25*(4), 298-313. https://doi.org/10.1177/1073858418805427

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D. J., Buttrick, N. R., Chartier, C. R., ... & Szecsi, P. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309-331. https://doi.org/10.1177%2F2515245920958687

Evans, R. I., Rozelle, R. M., Mittelmark, M. B., Hansen, W. B., Bane, A. L., & Havis, J. (1978). Deterring the onset of smoking in children: Knowledge of immediate physiological effects and coping with peer pressure, media pressure, and parent modeling 1. Journal of applied social psychology, 8(2), 126-135.

Fiedler, K., McCaughey, L., & Prager J. (2021). Quo vadis, methodology: The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science.* https://doi.org/10.1177/1745691620970602

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd.

Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management, 41*(2), 632-643. https://doi.org/10.1177%2F0149206314525208

Gibson, J.J. (1950). *The perception of the visual world*. Houghton Mifflin.

Gibson, J.J. (1966). The senses considered as perceptual systems. Houghton Mifflin.

Grossman, J., & Mackenzie, F. J. (2005). The randomized controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine, 48,* 516-534. https://doi.org/10.1353/pbm.2005.0092

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620921521

Hacking, I. (1992a). Statistical language, statistical truth and statistical reason: The self-authentication of a style of scientific reasoning. In E. McMullin (Ed.), *The social dimensions of science* (Vol. 3, pp. 130-157). University of Notre Dame Press.

Hacking, I. (1992b). The self-vindication of the laboratory sciences. In A. Pickering (Ed.), *Science as practice and culture* (pp. 29-64). University of Chicago Press.

Hacking, I. (2002). *Historical ontology*. Harvard University Press.

Hall, J. F. (1971). *Verbal learning and retention*. Lippincott.

Haynes, W. A., Tomczak, A., & Khatri, P. (2018). Gene annotation bias impedes biomedical research. *Scientific Reports*, *8*(1), 1-7. https://doi.org/10.1038/s41598-018-19333-x

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*(2), 85-110. https://doi.org/10.1257/jep.9.2.85

Heilmann, A., Mehay, A., Watt, R. G., Kelly, Y., Durrant, J. E., van Turnhout, J., & Gershoff, E. T. (2021). Physical punishment and child outcomes: a narrative review of prospective studies. *The Lancet*, 2021. https://doi.org/10.1016/S0140-6736(21)00582-1

Henrich, J., Heine, S., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466, 29. https://doi.org/10.1038/466029a

Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve inference from observational data. *American Journal of Public Health, 108*(5), 616-619. https://doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., & Robins, J. M. (2008). Observational studies analyzed like randomized experiments. An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology, 19*(6), 776-779. https://doi.org/10.1097/EDE.0b013e3181875e61

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2013). Randomized trials analyzed as observational studies. *Annals of Internal Medicine, 159*(8), 560-562. https://doi.org/10.7326/0003-4819-159-8-201310150-00709

Hernán M. A., & Robins J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.

Holland, P. W., & Rubin, D. B. (1988). Causal inference in retrospective studies. *Evaluation Review, 12*(3), 203-231. https://doi.org/10.1177%2F0193841X8801200301

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, Series A, *171*(2), 481-502. https://doi.org/10.1111/j.1467-985X.2007.00527.x

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature, 48*(2), 399-423. https://doi.org/10.1257/jel.48.2.399

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Jadad, A., & Enkin, M. (2007). *Randomized controlled trials: Questions, answers, and musings* (2nd ed.). Blackwell.

Joseph, D., Chan, M. Y., Heintzelman, S. J., Tay, L., Diener, E., & Scotney, V. S. (2020). The experimental manipulation of affect: A meta-analysis of affect induction procedures. *Psychological Bulletin*, *146*(4), 355-375. https://doi.org/10.1037/bul0000224

Kardon, G. (2018). Life in triplicate. *Science*, *359*(6381), 1222-1222. https://doi.org/10.1126/science.aat0954

Khosrowi, D. (2019). Extrapolation of causal effects—hopes, assumptions, and the extrapolator's circle. *Journal of Economic Methodology*, *26*(1), 45-58. https://doi.org/10.1080/1350178X.2018.1561078

Knorr-Cetina, K. (2009). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Krauss, A. (2018). Why all randomized controlled trials produce biased results. *Annals of Medicine, 50*(4), 312-322. https://doi.org/10.1080/07853890.2018.1453233

Kruglanski, A. W. (1975). The Human Subject in the Psychology Experiment: Fact and Artifact. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8, pp. 101-147). Academic Press.

Levitt, S., & List, J. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153-174. https://doi.org/10.1257/jep.21.2.153

Li, J., Zyphur, M. J., Sugihara, G., & Laub, P. J. (2021). Beyond linearity, stability, and equilibrium: The edm package for empirical dynamic modeling and convergent cross-mapping in Stata. *The Stata Journal*, *21*(1), 220-258. https://doi.org/10.1177/1536867X211000030

Little, T. D. (2019). Series Editor's note for the book C. S. Reichardt. (2019). *Quasi-Experimentation. A guide to design and analysis*. Guilford.

Lobo, R. (2017). Hormone-replacement therapy: Current thinking. *Nature Reviews Endocrinology, 13*, 220-231. https://doi.org/10.1038/nrendo.2016.164

Mackie, J. L. (1980). *The cement of the universe: A study of causation.* Clarendon Press.
MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Lawrence Erlbaum.
MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593-614. https://doi.org/10.1146/annurev.psych.58.110405
Maydeu-Olivares, A., Shi, D., & Fairchild, A. J. (2020). Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, *25*(2), 243-258. https://doi.org/10.1037/met0000226
McBeath, M. K., Shaffer, D. M., & Kaiser, M. K. (1995). How baseball outfielders determine where to run to catch fly balls. *Science*, *268*(5210), 569-573. https://doi.org/10.1126/science.7725104
McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist*, 25(2), 179-210. https://doi.org/10.1177/000276428102500205
McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, *73*, 99-105. https://doi.org/10.1080/00031305.2018.1505655
Möllenkamp, M., Zeppernick, M., & Schreyögg, J. (2019). The effectiveness of nudges in improving the self-management of patients with chronic diseases: A systematic literature review. *Health Policy*, *123*(12), 1199-1209. https://doi.org/10.1016/j.healthpol.2019.09.008
Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201-218. https:// https://doi.org/10.1207/s15366359mea0204_1
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387.
Morgan, G. (2006). *Images of organization*. Sage. https:// https://doi.org/10.1037/0003-066X.38.4.379
Muller, S. (2015). Interaction and external validity: Obstacles to the policy relevance of randomized evaluations. *World Bank Economic Review, 29*(1), 217-25. https://doi.org/10.1093/wber/lhv027
Neisser, U. (1967). *Cognitive psychology*. Prentice-Hall.
Neisser, U. (1976). Cognition and reality: Principles and implications of cognitive psychology. Freeman.
Northcott, R. (in press). *Science for a fragile world*. Oxford University Press.
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349* (6251), aac 4716. https://doi.org/10.1126/science.aac4716
Pearl, J. (2009). *Causality* (2nd ed). Cambridge University Press.
Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579-595. https://doi.org/10.1214/14-STS486
Pearl, J., Glymour, M., & Jewell, N. (2016). *Causal inference in statistics: A primer*. Wiley.
Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford.
Reiss, J. (2008). *Error in economics: Towards a more evidence-based methodology*. Routledge.
Rosenbaum, P. R. (2015). How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and its Applications*, *2*, 21-48. https://doi.org/10.1146/annurev-statistics-010814-020201
Ross, L. (2021). Causal concepts in biology: How pathways differ from mechanisms and why it matters. *British Journal for the Philosophy of Science*, *72.1*, 131-158. https://doi.org/10.1093/bjps/axy078

Rossi, J. S. (2013). Statistical power analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology. Volume 2: Research methods in psychology* (2nd ed., pp. 71-108). (I.B. Weiner, Editor-in-Chief). John Wiley & Sons.

Rothwell, P. M. (2005). External validity of randomized controlled trials: To whom do the results of the trial apply? *Lancet, 365*, 82-93. https://doi.org/10.1016/s0140-6736(04)17670-8

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*(1), 2-14. https://doi.org/10.1207%2FS15327957PSPR0501_1

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment non-compliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods, 19*, 317-333. https://doi.org/10.1037/met0000013

Salpeter, S. R., Walsh, J. M., Greyber, E., Ornisston, T. M., & Salpeter, E. D. (2004) Mortality associated with hormone replacement therapy in younger snd older women: A meta-analysis. *Journal of General Internal Medicine, 19*(7), 791-804. https://doi.org/10.1111/j.1525-1497.2004.30281.x

Schulze, J., West, S. G., Freudenstein, J.-P., Schäpers, P., Mussel, P., Eid, M., & Krumm, S. (in press). Hidden framings and hidden asymmetries in the measurement of personality: A combined lens-model and frame-of-reference perspective. *Journal of Personality*. https://doi.org/10.1111/jopy.12586

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515-530. https://doi.org/10.1037/0022-3514.51.3.515

Spirtes, P., Glymour,C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7,* 3-18. https://doi.org/10.1037/1082-989x.7.1.3

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for causal inference.* Houghton-Mifflin.

Shamsollahi, A., Zyphur, M. J., & Ozkok O. (in press). From short-run to long-run effects in cross-lagged panel models: An example using HRM and organizational performance. *Organizational Research Methods*. https://doi.org/10.1177%2F1094428121993228

Smith, A. B. (2006). The state of research on the effects of physical punishment. *Social Policy Journal of New Zealand*, 27, 114-127.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*(6), 845-851. https://doi.org/10.1037/0022-3514.89.6.845

St. Clair, T., Hallberg, K., Cook, T. D. (2016). The validity and precision of the comparative interrupted time-series design: Three within-study comparisons. *Journal of Educational and Behavioral Statistics*, 41(3), 269-299. https://doi.org/10.3102/1076998616636854

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie, 227*(4), 280-292. https:// https://doi.org/10.1027/2151-2604/a000385

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, *364*(6437), eaav7893. https://doi.org/10.1126/science.aav7893

Stroebe, W., and Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59-71. https://doi.org/10.1177%2F1745691613514450

Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496-500. https://doi.org/10.1126/science.1227079

Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P., & Khatri, P. (2017). Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Research*, *45*(1), e1. https://doi.org/10.1093/nar/gkw797

Syed, M. (2021). *It's a 2 X 2 design all the way down: Social psychology's over-reliance on experiments needlessly restricts diversity in the field.* Society for Personality and Social Psychology Annual Convention, February 5. https://doi.org/10.31234/osf.io/u89e2

Thelen, E., & Smith, L. B. (2007). Dynamic systems theories. In R. M. Lerner (Ed.), *Handbook of Child Psychology* (Vol. 1); *Theoretical Models of Human Development* (6th ed., pp. 258-310). Wiley.

Thoemmes, F. J., West, S. G., & Hill. E. (2009). Propensity score matching in a meta-analysis comparing randomized and non-randomized studies. *Multivariate Behavioral Research, 44*, 854 (abstract). https://doi.org/10.1080/00273170903467521

Tsonis, A. A., Deyle, E. R., May, R. M., Sugihara, G., Swanson, K., Verbeten, J. D., & Wang, G. (2015). Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences*. *112*(11), 3253-3256. https://doi.org/10.1073/pnas.1420291112

Ushio, M., Hsieh, C. H., Masuda, R., Deyle, E. R., Ye, H., Chang, C. W., ... & Kondoh, M. (2018). Fluctuating interaction network and time-varying stability of a natural fish community. *Nature*, *554*(7692), 360-363. https://doi.org/10.5281/zenodo.1039387

Vallania, F., Tam, A., Lofgren, S., Schaffert, S., Azad, T. D., Bongen, E., ... & Engleman, E. (2018). Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nature Communications*, *9*(1), 1-8. https://doi.org/10.1038/s41467-018-07242-6

Vanderweele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction.* Oxford University Press.

Van Nes, E. H., Scheffer, M., Brovkin, V., Lenton, T. M., Ye, H., Deyle, E., & Sugihara, G. (2015). Causal feedbacks in climate change. *Nature Climate Change*, *5*(5), 445-448. https://doi.org/10.1038/srep21691

Voelkle, M. C., Gische, C., Driver, C. C., & Lindenberger, U. (2018). The role of time in the quest for understanding psychological mechanisms. *Multivariate Behavioral Research, 53*(6), 782-805. https://doi.org/10.1080/00273171.2018.1496813

West, S. G., & Aiken, L. S. (1997). Towards understanding individual effects in multiple component prevention programs: Design and analysis strategies. In K. Bryant, M. Windle, & S. G. West (Eds.), *Recent advances in prevention methodology: Alcohol and substance abuse research* (pp. 167-209). American Psychological Association. https://doi.org/10.1037/10222-006

West, S. G., Cham, H., & Liu, Y. (2014). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Personality and Social Psychology* (2nd ed., pp. 49-80). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.007

West, S. G., Duan, N., Pequegnat, W., Gaist, P., DesJarlais, D., Holtgrave, D., … & Mullen, P. D. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health, 98*(8), 1359-1366. https://doi.org/10.2105/AJPH.2007.124446

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods, 15*, 18-37. https://doi.org/10.1037/a0015917

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health, 39,* 453-469. https://doi.org/10.1146/annurev-publhealth-040617-013507

Wittgenstein, L. (1958). *Philosophical investigations*. Macmillan.

Worrall, J. (2007). Why there's no cause to randomize. *British Journal for the Philosophy of Science, 58*(3), 451-488. https://doi.org/10.1093/bjps/axm024

Yarkoni, T. (2020). The generalizability crisis. *The Behavioral and Brain Sciences*, 1-37. https://doi.org/10.1017/S0140525X20001685

Ye, H., Deyle, E. R., Gilarranz, L. J., & Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports, 5*(1), 1-9. https://doi.org/10.1038/srep14750

Yeh, R. W., Valsdottir, L. R., Yeh, M. W., Shen, C., Kramer, D. B., Strom, J. B., ... & Nallamothu, B. K. (2018). Parachute use to prevent death and major trauma when jumping from aircraft: Randomized controlled trial. *BMJ, 363*, k5094. https://doi.org/10.1136/bmj.k5094

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics, 134*(2), 557-598. https://doi.org/10.1093/qje/qjy029

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., ... & Diener, E. (2020). From data to causes I: Building a general cross-lagged panel model (GCLM). *Organizational Research Methods, 23*(4), 651-687. https://doi.org/10.1177%2F1094428119847278