



# Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences

## Target Article

**Cite this article:** Almaatouq A, Griffiths TL, Suchow JW, Whiting ME, Evans J, Watts DJ. (2024) Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences* **47**, e33: 1–70. doi:10.1017/S0140525X22002874

Target Article Accepted: 27 November 2022  
Target Article Manuscript Online: 21 December 2022

Commentaries Accepted: 14 April 2023

### Keywords:

cumulative knowledge; experiments; generalizability; (in)commensurability

**What is Open Peer Commentary?** What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 18) and an Author's Response (p. 65). See [bbsonline.org](https://bbsonline.org) for more information.

### Corresponding author:

Abdullah Almaatouq;  
Email: [amaatouq@mit.edu](mailto:amaatouq@mit.edu)

Abdullah Almaatouq<sup>a</sup> , Thomas L. Griffiths<sup>b</sup>, Jordan W. Suchow<sup>c</sup>,  
Mark E. Whiting<sup>d</sup> , James Evans<sup>e,f</sup> and Duncan J. Watts<sup>g</sup>

<sup>a</sup>Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA; <sup>b</sup>Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA; <sup>c</sup>School of Business, Stevens Institute of Technology, Hoboken, NJ, USA; <sup>d</sup>School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA; <sup>e</sup>Department of Sociology, University of Chicago, Chicago, IL, USA; <sup>f</sup>Santa Fe Institute, Santa Fe, NM, USA and <sup>g</sup>Department of Computer and Information Science, Annenberg School of Communication, and Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA, USA  
[amaatouq@mit.edu](mailto:amaatouq@mit.edu); [tomg@princeton.edu](mailto:tomg@princeton.edu); [jws@stevens.edu](mailto:jws@stevens.edu); [markew@seas.upenn.edu](mailto:markew@seas.upenn.edu); [jevans@uchicago.edu](mailto:jevans@uchicago.edu); [djwatts@seas.upenn.edu](mailto:djwatts@seas.upenn.edu)

## Abstract

The dominant paradigm of experiments in the social and behavioral sciences views an experiment as a test of a theory, where the theory is assumed to generalize beyond the experiment's specific conditions. According to this view, which Alan Newell once characterized as “playing twenty questions with nature,” theory is advanced one experiment at a time, and the integration of disparate findings is assumed to happen via the scientific publishing process. In this article, we argue that the process of integration is at best inefficient, and at worst it does not, in fact, occur. We further show that the challenge of integration cannot be adequately addressed by recently proposed reforms that focus on the reliability and replicability of individual findings, nor simply by conducting more or larger experiments. Rather, the problem arises from the imprecise nature of social and behavioral theories and, consequently, a lack of commensurability across experiments conducted under different conditions. Therefore, researchers must fundamentally rethink how they design experiments and how the experiments relate to theory. We specifically describe an alternative framework, integrative experiment design, which intrinsically promotes commensurability and continuous integration of knowledge. In this paradigm, researchers explicitly map the design space of possible experiments associated with a given research question, embracing many potentially relevant theories rather than focusing on just one. Researchers then iteratively generate theories and test them with experiments explicitly sampled from the design space, allowing results to be integrated across experiments. Given recent methodological and technological developments, we conclude that this approach is feasible and would generate more-reliable, more-cumulative empirical and theoretical knowledge than the current paradigm – and with far greater efficiency.

## 1. Introduction

*You can't play 20 questions with Nature and win.* (Newell, 1973)

Fifty years ago, Allen Newell summed up the state of contemporary experimental psychology as follows: “Science advances by playing twenty questions with nature. The *proper* tactic is to frame a general question, hopefully binary, that can be attacked experimentally. Having settled that bits-worth, one can proceed to the next ... *Unfortunately, the questions never seem to be really answered, the strategy does not seem to work*” (italics added for emphasis).

The problem, Newell noted, was a lack of coherence among experimental findings. “We never seem in the experimental literature to put the results of all the experiments together,” he wrote, “Innumerable aspects of the situations are permitted to be suppressed. Thus, no way exists of knowing whether the earlier studies are in fact commensurate with whatever ones are under present scrutiny, or are in fact contradictory.” Referring to a collection of papers by prominent experimentalists, Newell concluded that although it was “exceedingly clear that each paper made a contribution ... I couldn't convince myself that it would add up, even in thirty more years of trying, even if one had another 300 papers of similar, excellent ilk.”

More than 20 years after Newell's imagined future date, his outlook seems, if anything, optimistic. To illustrate the problem, consider the phenomenon of group “synergy,” defined as the performance of an interacting group exceeding that of an equivalently sized “nominal group” of individuals working independently (Hill, 1982; Larson, 2013). A century of

experimental research in social psychology, organizational psychology, and organizational behavior has tested the performance implications of working in groups relative to working individually (Allen & Hecht, 2004; Richard Hackman & Morris, 1975; Husband, 1940; Schulz-Hardt & Mojzisch, 2012; Tasca, 2021; Watson, 1928), but substantial contributions can also be found in cognitive science, communications, sociology, education, computer science, and complexity science (Allport, 1924; Arrow, McGrath, & Berdahl, 2000; Barron, 2003; Devine, Clayton, Dunford, Searing, & Pryce, 2001). In spite of this attention across time and disciplines – or maybe because of it – this body of research often reaches inconsistent or conflicting conclusions. For example, some studies find that interacting groups outperform

individuals because they are able to distribute effort (Laughlin, Bonner, & Miner, 2002), share information about high-quality solutions (Mason & Watts, 2012), or correct errors (Mao, Mason, Suri, & Watts, 2016), whereas other studies find that “process losses” – including social loafing (Harkins, 1987; Karau & Williams, 1993), groupthink (Janis, 1972), and interpersonal conflict (Steiner, 1972) – cause groups to underperform their members.

As we will argue, the problem is not that researchers lack theoretically informed hypotheses about the causes and predictors of group synergy; to the contrary, the literature contains dozens, or possibly even hundreds, of such hypotheses. Rather, the problem is that because each of these experiments was designed with the goal of testing a hypothesis but, critically, *not* with the goal of explicitly comparing the results with other experiments of the same general class, researchers in this space have no way to articulate how similar or different their experiment is from anyone else's. As a result, it is impossible to determine – via systematic review, meta-analysis, or any other ex-post method of synthesis – how all of the potentially relevant factors jointly determine group synergy or how their relative importance and interactions change over contexts and populations.

Nor is group synergy the only topic in the social and behavioral sciences for which one can find a proliferation of irreconcilable theories and empirical results. For any substantive area of the social and behavioral sciences on which we have undertaken a significant amount of reading, we see hundreds of experiments that each tests the effects of some independent variables on other dependent variables while suppressing innumerable “aspects of the situation.”<sup>1</sup> Setting aside the much-discussed problems of replicability and reproducibility, many of these papers are interesting when read in isolation, but it is no more possible to “put them all together” today than it was in Newell's time (Almaatouq, 2019; Muthukrishna & Henrich, 2019; Watts, 2017).

Naturally, our subjective experience of reading across several domains of interest does not constitute proof that successful integration of many independently designed and conducted experiments cannot occur in principle, or even that it has not occurred in practice. Indeed it is possible to think of isolated examples, such as mechanism design applied to auctions (Myerson, 1981; Vickrey, 1961) and matching markets (Aumann & Hart, 1992; Gale & Shapley, 1962), in which theory and experiment appear to have accumulated into a reasonably self-consistent, empirically validated, and practically useful body of knowledge. We believe, however, that these examples represent rare exceptions and that examples such as group synergy are far more typical.

We propose two explanations for why not much has changed since Newell's time. The first is that *not everyone agrees with the premise of Newell's critique* – that “putting things together” is a pressing concern for the scientific enterprise. In effect, this view holds that the approach Newell critiqued (and that remains predominant in the social and behavioral sciences) is sufficient for accumulating knowledge. Such accumulation manifests itself indirectly through the scientific publishing process, with each new paper building upon earlier work, and directly through literature reviews and meta-analyses. The second explanation for the lack of change since Newell's time is that even if one accepts Newell's premise, *neither Newell nor anyone else has proposed a workable alternative*; hence, the current paradigm persists by default in spite of its flaws.<sup>2</sup>

In the remainder of this paper, we offer our responses to the two explanations just proposed. Section 2 addresses the first explanation, describing what we call the “one-at-a-time”

ABDULLAH ALMAATOUQ is the Douglas Drane Career professor of Information Technology at the Massachusetts Institute of Technology. Abdullah has two primary research interests: The first explores how groups, organizations, and societies can optimize their decision making. The second focuses on improving social and behavioral research methodology. He holds a PhD in Computational Science and Engineering from MIT.

THOMAS L. GRIFFITHS is the Henry R. Luce professor of Information Technology, Consciousness, and Culture at Princeton University. His research explores connections between psychology and computer science, using ideas from machine learning and artificial intelligence to understand how people solve the challenging computational problems they encounter in everyday life.

JORDAN W. SUCHOW is an assistant professor at Stevens Institute of Technology. Suchow's research sits at the intersection of cognitive science and information systems, by studying the cognitive underpinnings of sociotechnical systems and leveraging information technologies such as crowdsourcing platforms to scale experimental work in cognitive science. He holds a B.S. in Computer Science from Brandeis University and an A.M. and PhD in Psychology from Harvard University.

MARK E. WHITING is a senior computational social scientist at the CSSLab at the University of Pennsylvania with affiliations in Computer & Information Science and Operations, Information, and Decisions. His research involves designing empirical systems to study how people behave and coordinate at scale. He holds bachelor's and master's degrees in Industrial Design from Royal Melbourne Institute of Technology and Korea Advanced Institute of Science and Technology, respectively, and a PhD in Mechanical Engineering from Carnegie Mellon University.

JAMES EVANS is the Max Palevsky professor of Sociology, Director of Knowledge Lab and the Computational Social Science Program at the University of Chicago, and external faculty at the Santa Fe Institute. His research explores and seeks to understand the nature of collective thinking and knowing – from aggregate imagination to shared certainty – using large-scale data, generative modeling, and adaptive experiments. He holds a PhD in Sociology from Stanford University.

DUNCAN J. WATTS is the Stevens University professor and 23rd Penn Integrates Knowledge (PIK) professor and Director of the CSSLab at the University of Pennsylvania, where he holds faculty appointments in the Department of Computer and Information Science, The Annenberg School of Communications, and the Operations, Information, and Decisions Department in the Wharton School. He holds a BSc in Physics from the University of New South Wales and a PhD in Theoretical and Applied Mechanics from Cornell University.

paradigm and arguing that it is poorly suited to the purpose of integrating knowledge over many studies in large part because it was not designed for that purpose. We also argue that existing mechanisms for integrating knowledge, such as systematic reviews and meta-analyses, are insufficient on the grounds that they, in effect, assume commensurability. If the studies that these methods are attempting to integrate cannot be compared with one another, because they were not designed to be commensurable, then there is little that ex-post methods can do.<sup>3</sup> Rather, an alternative approach to designing experiments and evaluating theories is needed. Section 3 addresses the second explanation by describing such an alternative, which we call the “integrative” approach, that is explicitly designed to integrate knowledge about a particular problem domain. Although integrative experiments of the sort we describe may not have been possible in Newell’s day, we argue that they can now be productively pursued in parts of the social and behavioral sciences thanks to increasing theoretical maturity and methodological developments. To illustrate this point, section 4 illustrates the potential of the integrative approach by describing three experiments that are first steps in its direction. Finally, section 5 outlines questions and concerns we have encountered and offers our response.

## 2. The “one-at-a-time” paradigm

In the simplest version of what we call the “one-at-a-time” approach to experimentation, a researcher poses a question about the relation between one independent and one dependent variable and then offers a theory-motivated hypothesis that the relation is positive or negative. Next, the researcher devises an experiment to test this hypothesis by introducing variability in the independent variable, aiming to reject the “null hypothesis” that the proposed dependency does not exist on the basis of the evidence, quantified by a  $p$ -value. If the null hypothesis is successfully rejected, the researcher concludes that the experiment corroborates the theory and then elaborates on potential implications, both for other experiments and for phenomena outside the lab.

In practice, one-at-a-time experiments can be considerably more complex. The researcher may articulate hypotheses about more than one independent variable, more than one dependent variable, or both. The test itself may focus on effect sizes or confidence intervals rather than statistical significance, or it may compare two or more competing hypotheses. Alternatively, both the hypothesis and the test may be qualitative in nature. Regardless, each experiment tests at most a small number of theoretically informed hypotheses in isolation by varying at most a small number of parameters. By design, all other factors are held constant. For example, a study of the effect of reward or punishment on levels of cooperation typically focuses on the manipulation of theoretical interest (e.g., introducing a punishment stage between contribution rounds in a repeated game) while holding fixed other parameters, such as the numerical values of the payoffs or the game’s length (Fehr & Gächter, 2000). Similarly, a study of the effect of network structure on group performance typically focuses on some manipulation of the underlying network while holding fixed the group size or the time allotted to perform the task (Almaatouq et al., 2020; Becker, Brackbill, & Centola, 2017).

### 2.1. The problem with the one-at-a-time paradigm

As Newell himself noted, this approach to experimentation seems reasonable. After all, the sequence of *question* → *theory* →

*hypothesis* → *experiment* → *analysis* → *revision to theory* → *repeat* appears to be almost interchangeable with the scientific method itself. Nonetheless, the one-at-a-time paradigm rests on an important but rarely articulated assumption: That because the researcher’s purpose in designing an experiment is to test a theory of interest, the only constructs of interest are those that the theory itself explicitly articulates as relevant. Conversely, where the theory is silent, the corresponding parameters are deemed to be irrelevant. According to this logic, articulating a precise theory leads naturally to a well-specified experiment with only one, or at most a few, constructs in need of consideration. Correspondingly, theory can aid the interpretation of the experiment’s results – and can be generalized to other cases (Mook, 1983; Zelditch, 1969).

Unfortunately, while such an assumption may be reasonable in fields such as physics, it is rarely justified in the social and behavioral sciences (Debrouwere & Rosseel, 2022; Meehl, 1967). Social and behavioral phenomena exhibit higher “causal density” (or what Meehl called the “crud factor”) than physical phenomena, such that the number of potential causes of variation in any outcome is much larger than in physics and the interactions among these causes are often consequential (Manzi, 2012; Meehl, 1990b). In other words, the human world is vastly more complex than the physical one, and researchers should be neither surprised nor embarrassed that their theories about it are correspondingly less precise and predictive (Watts, 2011). The result is that theories in the social and behavioral sciences are rarely articulated with enough precision or supported by enough evidence for researchers to be sure which parameters are relevant and which can be safely ignored (Berkman & Wilson, 2021; Meehl, 1990b; Turner & Smaldino, 2022; Yarkoni, 2022). Researchers working independently in the same domain of inquiry will therefore invariably make design choices (e.g., parameter settings, subject pools) differently (Brezna et al., 2022; Gelman & Loken, 2014). Moreover, because the one-at-a-time paradigm is premised on the (typically unstated) assumption that theories dictate the design of experiments, the process of making design decisions about constructs that are not specified under the theory being tested is often arbitrary, vague, undocumented, or (as Newell puts it) “suppressed.”

### 2.2. The universe of possible experiments

To express the problem more precisely, it is useful to think of a one-at-a-time experiment as a sample from an implicit universe of possible experiments in a domain of inquiry. Before proceeding, we emphasize that neither the sample nor the universe is typically acknowledged in the one-at-a-time paradigm. Indeed, it is precisely the transition from implicit to explicit construction of the sampling universe that forms the basis of the solution we describe in the next section.

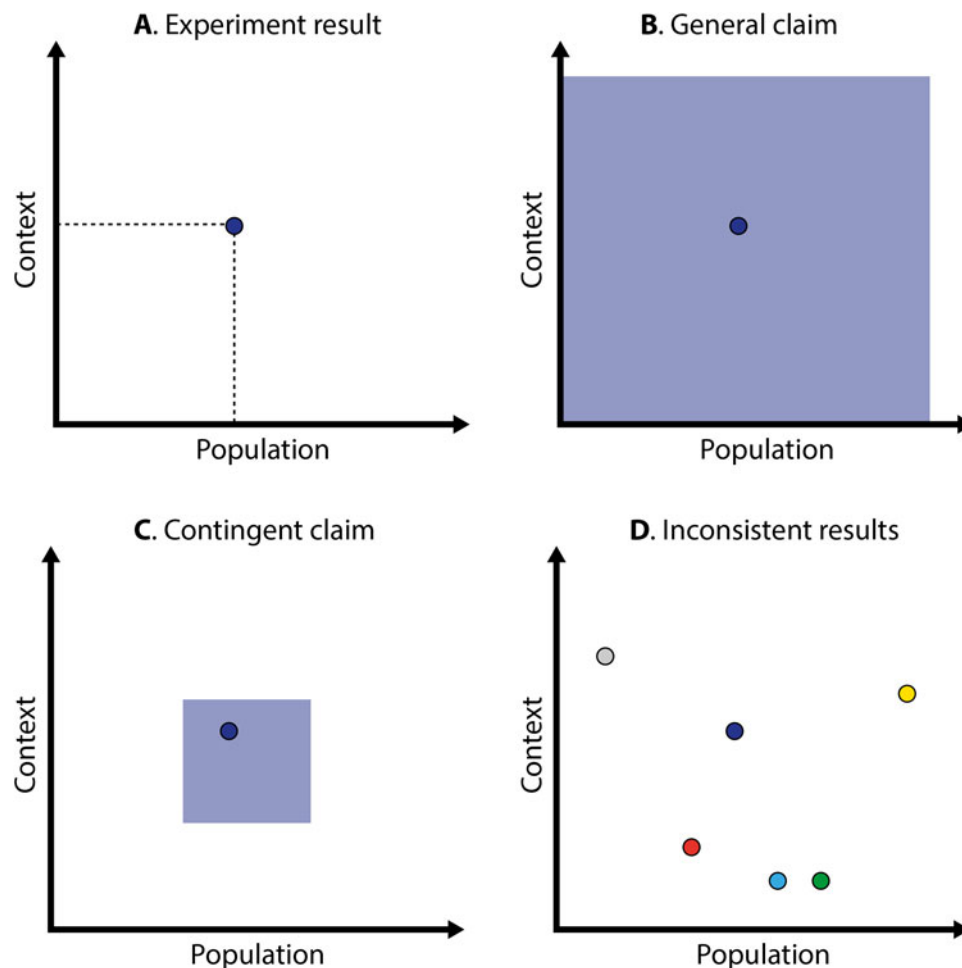
In imagining such a universe, it is useful to distinguish the independent variables needed to define the effect of interest – the experimental manipulation – from the experiment’s *context*. We define this context as the set of independent variables that are hypothesized to moderate the effect in question as well as the nuisance parameters (which, strictly speaking, are also independent variables) over which the effect is expected to generalize and that correspond to the design choices the researcher makes about the specific experiment that will be conducted. For example, an experiment comparing the performance of teams to that of individuals not only will randomize participants into a set of experimental conditions (e.g., individuals vs. teams of varying

sizes), but will also reflect decisions about other contextual features, including, for example, the specific tasks on which to compare performance, where each task could then be parameterized along multiple dimensions (Almaatouq, Alsobay, Yin, & Watts, 2021a; Larson, 2013). Other contextual choices include the incentives provided to participants, time allotted to perform the task, modality of response, and so on. Similarly, we define the *population* of the experiment as a set of measurable attributes that characterize the sample of participants (e.g., undergraduate women in the United States aged 18–23 with a certain distribution of cognitive reflection test scores). Putting all these choices together, we can now define an abstract space of possible experiments, the dimensions of which are the union of the context and population. We call this space the *design space* on the grounds that every conceivable design of the experiment is describable by some choice of parameters that maps to a unique point in the space.<sup>4</sup> (Although this is an abstract way of defining what we mean by the experiment design space, we will suggest concrete and practical ways of defining it later in the article.)

Figure 1 shows a simplified rendering of a design space and illustrates several important properties of the one-at-a-time paradigm. Figure 1A shows a single experiment conducted in a particular context with a particular sample population. The color of the

point represents the “result” of the experiment: The effect of one or more independent variables on some dependent variable. In the absence of a theory, nothing can be concluded from the experiment alone, other than that the observed result holds for one particular sample of participants under one particular context. From this observation, the appeal of strong theory becomes clear: By framing an experiment as a test of a theory, rather than as a measurement of the relationship between dependent and independent variables (Koyré, 1953), the observed results can be generalized well beyond the point in question, as shown in Figure 1B. For example, while a methods section of an experimental paper might note that the participants were recruited from the subject pool at a particular university, it is not uncommon for research articles to report findings as if they apply to all of humanity (Henrich, Heine, & Norenzayan, 2010). According to this view, theories (and in fields such as experimental economics, formal models) are what help us understand the world, whereas experiments are merely instruments that enable researchers to test theories (Lakens, Uygun Tunç, & Necip Tunç, 2022; Levitt & List, 2007; Mook, 1983; Zelditch, 1969).

As noted above, however, we rarely expect theories in the social and behavioral sciences to be universally valid. The ability of the theory in question to generalize the result is therefore almost



**Figure 1.** Implicit design space. Panel A depicts a single experiment (a single point) that generates a result in a particular sample population and context; the point's color represents a relationship between variables. Panel B depicts the expectation that results will generalize over broader regions of conditions. Panel C shows a result that applies to a bounded range of conditions. Panel D illustrates how isolated studies about specific hypotheses can reach inconsistent conclusions, as represented by different-colored points.



always limited to some region of the design space that includes the sampled point but not the entire space, as shown in [Figure 1C](#). While we expect that most researchers would acknowledge that they lack evidence for unconstrained generality over the population, it is important to note that there is nothing special about the subjects. In principle, what goes for subjects also holds for contexts (Simons, Shoda, & Lindsay, 2017; Yarkoni, 2022). Indeed, as Brunswik long ago observed, "...proper sampling of situations and problems may in the end be more important than proper sampling of subjects, considering the fact that individuals are probably on the whole much more alike than are situations among one another" (Brunswik, 1947).

Unfortunately, because the design space is never explicitly constructed, and hence the sampled point has no well-defined location in the space, the one-at-a-time paradigm cannot specify a proposed domain of generalizability. Instead, any statements regarding "scope" or "boundary" conditions for a finding are often implicit and qualitative in nature, leaving readers to assume the broadest possible generalizations. These scope conditions may appear in an article's discussion section but typically not in its title, abstract, or introduction. Rarely, if ever, is it possible to precisely identify, based on the theory alone, over what domain of the design space one should expect an empirical result to hold (Cesario, 2014, 2022).

### 2.3. Incommensurability leads to irreconcilability

Given that the choices about the design of experiments are not systematically documented, it becomes impossible to establish how similar or different two experiments are. This form of incommensurability, whereby experiments about the same effect of interest are incomparable, generates a pattern like that shown in [Figure 1D](#), where inconsistent and contradictory findings appear in no particular order or pattern (Levinthal & Rosenkopf, 2021). If one had a metatheory that specified precisely under what conditions (i.e., over what region of parameter values in the design space) each theory should apply, it might be possible to reconcile the results under that metatheory's umbrella, but rarely do such metatheories exist (Muthukrishna & Henrich, 2019). As a result, the one-at-a-time paradigm provides no mechanism by which to determine whether the observed differences (a) are to be expected on the grounds that they lie in distinct subdomains governed by different theories, (b) represent a true disagreement between competing theories that make different claims on the same subdomain, or (c) indicate that one or both results are likely to be wrong and therefore require further replication and scrutiny. In other words, inconsistent findings arising in the research literature are essentially irreconcilable (Almaatouq, 2019; Muthukrishna & Henrich, 2019; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016; Watts, 2017; Yarkoni, 2022).

Critically, the absence of commensurability also creates serious problems for existing methods of synthesizing knowledge such as systematic reviews and meta-analyses. As all these methods are post-hoc, meaning that they are applied after the studies in question have been completed, they are necessarily reliant on the designs of the experiments they are attempting to integrate. If those designs do not satisfy the property of commensurability (again, because they were never intended to), then ex-post methods are intrinsically limited in how much they can say about observed differences. A concrete illustration of this problem has emerged recently in the context of "nudging" due to the publication of a large meta-analysis of over 400 studies spanning a wide

range of contexts and interventions (Mertens, Herberz, Hahnel, & Brosch, 2022). The paper was subsequently criticized for failing to account adequately for publication bias (Maier et al., 2022), the quality of the included studies (Simonsohn, Simmons, & Nelson, 2022), and their heterogeneity (Szasz et al., 2022). While the first two of these problems can be addressed by proposed reforms in science, such as universal registries of study designs (which are designed to mitigate publication bias) and adoption of preanalysis plans (which are specified to improve study quality), the problem of heterogeneity requires a framework for expressing study characteristics in a way that is commensurate. If two studies are different, that is, a meta-analysis is left with no means to incorporate information from both of them that properly accounts for their differences. Thus, while meta-analyses (and reviews more generally) can acknowledge the importance of moderating variables, they are inherently limited in their ability to do so by the commensurability of the underlying studies.

Finally, we note that the lack of commensurability is also unaddressed by existing proposals to improve the reliability of science by, for example, increasing sample sizes, calculating effect sizes rather than measures of statistical significance, replicating findings, or requiring preregistered designs. Although these practices can indeed improve the reliability of individual findings, they are not concerned directly with the issue of how many such findings "fit together" and hence do not address our fundamental concern with the one-at-a-time framework. In other words, just as Newell claimed 50 years ago, improving the commensurability of experiments – and the theories they seek to test – will require a paradigmatic shift in how we think about experimental design.

## 3. From one-at-a-time to integrative by design

We earlier noted that a second explanation for the persistence of the one-at-a-time approach is the lack of any realistic alternative. Even if one sees the need for a "paradigmatic shift in how we think about experimental design," it remains unclear what that shift would look like and how to implement it. To address this issue, we now describe an alternative approach, which we call "integrative" experimentation, that can resolve some of the difficulties described previously. In general terms, the one-at-a-time approach starts with a single, often very specific, theoretically informed hypothesis. In contrast, the integrative approach starts from the position of embracing many potentially relevant theories: All sources of measurable experimental-design variation are potentially relevant, and decisions about which parameters are relatively more or less important are to be answered empirically. The integrative approach proceeds in three phases: (1) Constructing a design space, (2) sampling from the design space, and (3) building theories from the resulting data. The rest of this section elucidates these three main conceptual components of the integrative approach.

### 3.1. Constructing the design space

The integrative approach starts by explicitly constructing the design space. Experiments that have already been conducted can then be assigned well-defined coordinates, whereas those not yet conducted can be identified as as-yet-unsampled points. Critically, the differences between any pair of experiments that share the same effect of interest – whether past or future – can be determined; thus, it is possible to precisely identify the similarities and differences between two designs. In other words, commensurability is "baked in" by design.

How should the design space be constructed in practice? The method will depend on the domain of interest but is likely to entail a discovery stage that identifies candidate dimensions from the literature. Best practices for constructing the design space will emerge with experience, giving birth to a new field of what we tentatively label “research cartography”: The systematic process of mapping out research fields in design spaces. Efforts in research cartography are likely to benefit from and contribute to ongoing endeavors to produce formal ontologies in social and behavioral science research and other disciplines, in support of a more integrative science (Larson & Martone, 2009; Rubin et al., 2006; Turner & Laird, 2012).

To illustrate this process, consider the phenomenon of group synergy discussed earlier. Given existing theory and decades of experiments, one might expect the existence and strength of group synergy to depend on the task: For some tasks, interacting groups might outperform nominal groups, whereas for others, the reverse might hold. In addition, synergy might (or might not) be expected depending on the specific composition of the group: Some combinations of skills and other individual attributes might lead to synergistic performance; other combinations might not. Finally, group synergy might depend on “group processes,” defined as variables such as the communications technology or incentive structure that affect how group members interact with one another, but which are distinct both from the individuals themselves and their collective task.

Given these three broad sources of variation, an integrative approach would start by identifying the dimensions associated with each, as suggested either by prior research or some other source of insight such as practical experience. In this respect, research cartography resembles the process of identifying the nodes of a nomological network (Cronbach & Meehl, 1955; Preckel & Brunner, 2017) or the dimensions of methodological diversity for a meta-analysis (Higgins, Thompson, Deeks, & Altman, 2003); however, it will typically involve many more dimensions and require the “cartographer” to assign numerical coordinates to each “location” in the space. For example, the literature on group performance has produced several well-known task taxonomies, such as those by Shaw (1963), Hackman (1968), Steiner (1972), McGrath (1984), and Wood (1986). Task-related dimensions of variation (e.g., divisibility, complexity, solution demonstrability, and solution multiplicity) would be extracted from these taxonomies and used to label tasks that have appeared in experimental studies of group performance. Similarly, prior work has variously suggested that group performance depends on the composition of the group with respect to individual-level traits as captured by, say, average skill (Bell, 2007; Devine & Philips, 2001; LePine, 2003; Stewart, 2006), skill diversity (Hong & Page, 2004; Page, 2008), gender diversity (Schneid, Isidor, Li, & Kabst, 2015), social perceptiveness (Engel, Woolley, Jing, Chabris, & Malone, 2014; Kim et al., 2017; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010), and cognitive-style diversity (Aggarwal & Woolley, 2018; Ellemers & Rink, 2016), all of which could be represented as dimensions of the design space. Finally, group-process variables might include group size (Mao et al., 2016), properties of the communication network (Almaatouq, Rahimian, Burton, & Alhajri, 2022; Becker et al., 2017; Mason & Watts, 2012), and the ability of groups to reorganize themselves (Almaatouq et al., 2020). Together, these variables might identify upward of 50 dimensions that define a design space of possible experiments for studying group synergy through integrative experiment design,

where any given study should, in principle, be assignable to one unique point in the space.<sup>5</sup>

As this example illustrates, the list of possibly relevant variables can be long, and the dimensionality of the design space can therefore be large. Complicating matters, we do not necessarily know up front which of the many variables are in fact relevant to the effects of interest. In the example of group synergy, for instance, even an exhaustive reading of the relevant literature is not guaranteed to reveal all the ways in which tasks, groups, and group processes can vary in ways that meaningfully affect synergy. Conversely, there is no guarantee that all, or even most, of the dimensions chosen to represent the design space will play any important role in generating synergy. As a result, experiments that map to the same point in the design space could yield different results (because some important dimension is missing from the representation of the space), while in other cases, experiments that map to very different points yield indistinguishable behavior (because the dimensions along which they differ are irrelevant).

Factors such as these complicate matters in practice but do not present a fundamental problem to the approach described here. The integrative approach does not require the initial configuration of the space to be correct or its dimensionality to be fixed. Rather, the dimensionality of the space can be learned in parallel with theory construction and testing. Really, *the only critical requirement for constructing the design space is to do it explicitly and systematically by identifying potentially relevant dimensions* (either from the literature or from experience, including any known experiments that have already been performed) and by assigning coordinates to individual experiments along all identified dimensions. Using this process of explicit, systematic mapping of research designs to points in the design space (research cartography), the integrative approach ensures commensurability. We next will describe how the approach leverages commensurability to produce integrated knowledge in two steps: Via sampling, and via theory construction and testing.

### 3.2. Sampling from the design space

An important practical challenge to integrative experiment design is that the size of the design space (i.e., the number of possible experiments) increases exponentially with the number of identified dimensions  $D$ . To illustrate, assume that each dimension can be represented as a binary variable (0, 1), such that a given experiment either exhibits the property encoded in the dimension or does not. The number of possible experiments is then  $2^D$ . When  $D$  is reasonably small and experiments are inexpensive to run, it may be possible to exhaustively explore the space by conducting every experiment in a full factorial design. For example, when  $D = 8$ , there are 256 experiments in the design space, a number that is beyond the scale of most studies in the social and behavioral sciences but is potentially achievable with recent innovations in crowdsourcing and other “high-throughput” methods, especially if distributed among a consortium of labs (Byers-Heinlein et al., 2020; Jones et al., 2021). Moreover, running all possible experiments may not be necessary: If the goal is to estimate the impact that each variable has, together with their interactions, a random (or more efficient) sample of the experiments can be run (Auspurg & Hinz, 2014). This sample could also favor areas where prior work suggests meaningful variation will be observed. Using these methods, together with large samples, it is possible to run studies for higher values of  $D$  (e.g., 20). Section 4 describes examples of such studies.

Exhaustive and random sampling are both desirable because they allow unbiased evaluation of hypotheses that are not tethered to the experimental design – there is no risk of looking only at regions of the space that current hypotheses favor (Dubova, Moskvichev, & Zollman, 2022), and no need to collect more data from the design space because the hypotheses under consideration change. But as the dimensionality increases, exhaustive and random sampling quickly becomes infeasible. When  $D$  is greater than 20, the number of experiment designs grows to over 1 million, and when  $D = 30$ , it is over 1 billion. Given that the dimensionality of design spaces for even moderately complex problems could easily exceed these numbers, and that many dimensions will be not binary but ternary or greater, integrative experiments will require using different sampling methods.

Fortunately, there already exist a number of methods that enable researchers to efficiently sample high-dimensional design spaces (Atkinson & Donev, 1992; McClelland, 1997; Smucker, Krzywinski, & Altman, 2018; Thompson, 1933). For example, one contemporary class of methods is “active learning,” an umbrella term for sequential optimal experimental-design strategies that iteratively select the most informative design points to sample.<sup>6</sup> Active learning has become an important tool in the design of A/B tests in industry (Letham, Karrer, Ottoni, & Bakshy, 2019) and, more recently, of behavioral experiments in the lab (Baliatti, Klein, & Riedl, 2021).<sup>7</sup> Most commonly, an active learning process begins by conducting a small number of randomly selected experiments (i.e., points in the design space) and fitting a *surrogate model* to the outcome of these experiments. As we later elucidate, one can think of the surrogate model as a “theory” that predicts the outcome of all experiments in the design space, including those that have not been conducted. Then, a *sampling strategy* (also called an “acquisition function,” “query algorithm,” or “utility measure”) selects a new batch of experiments to be conducted according to the value of potential experiments. Notably, the choice of a surrogate model and sampling strategy is flexible, and the best alternative to choose will depend on the problem (Eyke, Koscher, & Jensen, 2021).<sup>8</sup>

We will not explore the details of these methods or their implementation,<sup>9</sup> as this large topic has been – and continues to be – extensively developed in the machine-learning and statistics communities.<sup>10</sup> For the purpose of our argument, it is

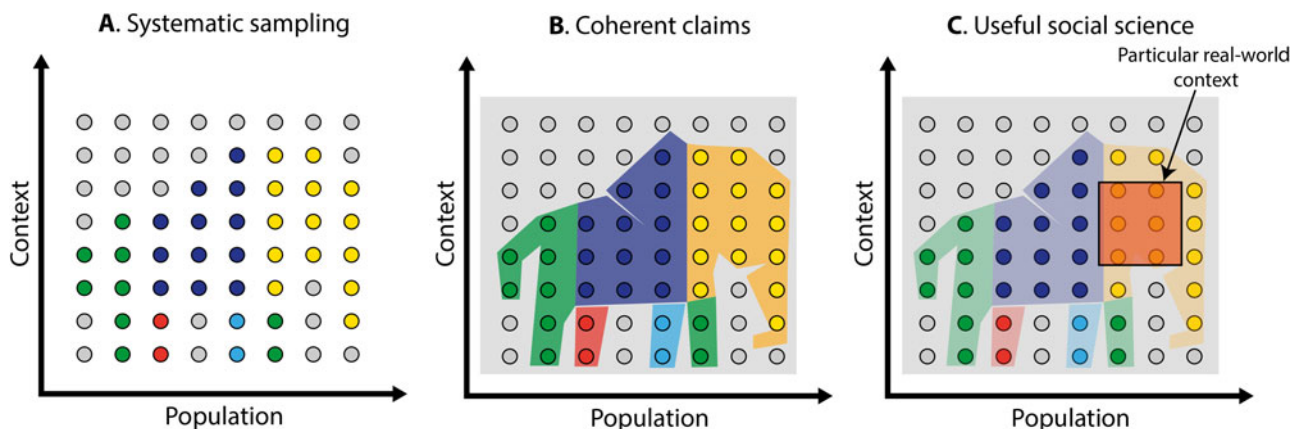
necessary only to convey that systematic sampling from the design space allows for unbiased evaluation of hypotheses (see Fig. 2A) and can leverage a relatively small number of sampled points in the design space to make predictions about every point in the space, the vast majority of which are never sampled (see Fig. 2B). Even so, by iteratively evaluating the model against newly sampled points and updating it accordingly, the model can learn about the entire space, including which dimensions are informative. As we explain next, this iterative process will also form the basis of theory construction and evaluation.

### 3.3. Building and testing theories

Much like in the one-at-a-time paradigm, the ultimate goal of integrative experiment design is to develop a reliable, cohesive, and cumulative theoretical understanding. However, because the integrative approach constructs and tests theories differently, the theories that tend to emerge from it depart from the traditional notion of theory in two regards. First, the shift to integrative experiments will change our expectations about what theories look like (Watts, 2014, 2017), requiring researchers to focus less on proposing novel theories that seek to differentiate themselves from existing theories by identifying new variables and their effects, and more on identifying theory boundaries, which may involve many known variables working together in complex ways. Second, although traditional theory development distinguishes sharply between basic and applied research, integrative theories will lend themselves to a “use-inspired” approach in which basic and applied science are treated as complements rather than as substitutes where one necessarily drives out the other (Stokes, 1997; Watts, 2017). We now describe each of these adaptations in more detail.

#### 3.3.1. Integrating and reconciling existing theories

As researchers sample experiments that cover more of the design space, simple theories and models that explain behavior with singular factors will no longer be adequate because they will fail to generalize. From a statistical perspective, the “bias-variance trade-off” principle identifies two ways a model (or theory) can fail to generalize: It can be too simple and thus unable to capture trends in the observed data, or too complex, overfitting the observed data



**Figure 2.** Explicit design space. Panel A shows that systematically sampling the space of possible experiments can reveal contingencies, thereby increasing the integrativeness of theories (as shown in panel B). Panel C depicts that what matters most is the overlap between the most practically useful conditions and domains defined by theoretical boundaries. The elephants in panels B and C represent the bigger picture that findings from a large number of experiments allow researchers to discern, but which is invisible to those from situated theoretical and empirical positions.



and manifesting great variance across datasets (Geman, Bienenstock, & Doursat, 1992). However, this variance decreases as the datasets increase in size and breadth, making oversimplification and reliance on personal intuitions more-likely causes of poor generalization. As a consequence, we must develop new kinds of theories – or metatheories – that capture the complexity of human behaviors while retaining the interpretability of simpler theories.<sup>11</sup> In particular, such theories must account for variation in behavior across the entire design space and will be subject to different evaluation criteria than those traditionally used in the social and behavioral sciences.

One such criterion is the requirement that theories generate “risky” predictions, defined roughly as quantitative predictions about as-yet unseen outcomes (Meehl, 1990b; Yarkoni, 2022). For example, in the “active sampling” approach outlined above, the surrogate model encodes prior theory and experimental results into a formal representation that (a) can be viewed as an explanation of all previously sampled experimental results and (b) can be queried for predictions treated as hypotheses. This dual status of the surrogate model as both explanation and prediction (Hofman et al., 2021; Nemesure, Heinz, Huang, & Jacobson, 2021; Yarkoni & Westfall, 2017) distinguishes it from the traditional notion of hypothesis testing. Rather than evaluating a theory based on how well it fits existing (i.e., in-sample) experimental data, the surrogate model is continually evaluated on its ability to predict new (i.e., out-of-sample) experimental data. Moreover, once the new data have been observed, the model is updated to reflect the new information, and new predictions are generated.

We emphasize that the surrogate model from the active learning approach is just one way to generate, test, and learn from risky predictions. Many other approaches also satisfy this criterion. For example, one might train a machine-learning model other than the surrogate model to estimate heterogeneity of treatment effects and to discover complex structures that were not specified in advance (Wager & Athey, 2018). Alternatively, one could use an interpretable, mechanistic, model. The only essential requirements for an integrative model are that it leverages the commensurability of the design space to in some way (a) *accurately explain* data that researchers have already observed, (b) *make predictions* about as-yet-unseen experiments, and then, having run those experiments, and (c) *integrate* the newly learned information to improve the model. If accurate predictions are achievable across some broad domain of the design space, the model can then be interpreted as supporting or rejecting various theoretical claims in a context-population-dependent way, as illustrated schematically in Figure 2B. Reflecting Merton’s (1968) call for “theories of the middle range,” a successful metatheory could identify the boundaries between empirically distinct regions of the design space (i.e., regions where different observed answers to the same research question pertain), making it possible to precisely state under what conditions (i.e., for which ranges of parameter values) one should expect different theoretically informed results to apply.

If accurate predictions are unachievable even after an arduous search, the result is not a failure of the integrative framework. Rather, it would be an example of the framework’s revealing a fundamental limit to prediction and, hence, explanation (Hofman, Sharma, & Watts, 2017; Martin, Hofman, Sharma, Anderson, & Watts, 2016; Watts et al., 2018).<sup>12</sup> In the extreme, when no point in the space is informative of any other point, generalizations of any sort are unwarranted. In such a scenario, applied research might still be possible, for example, by sampling the precise point of interest (Manzi, 2012), but the researcher’s drive to attain a

generalizable theoretical understanding of a domain of inquiry would be exposed as fruitless. Such an outcome would be disappointing, but from a larger scientific perspective, it is better to know what cannot be known than to believe in false promises. Naturally, whether such outcomes arise – and if so, how frequently – is itself an empirical question that the proposed framework could inform. With sufficient integrative experiments over many domains, the framework might yield a “meta-metatheory” that clarifies under which conditions one should (or should not) expect to find predictively accurate metatheories.

### 3.3.2. Bridging scientific and pragmatic knowledge

Another feature of integrative theories is that they will lend themselves to a “use-inspired” approach. Practitioners and researchers alike generally acknowledge that no single intervention, however evidence-based, benefits all individuals in all circumstances (i.e., across *populations* and *contexts*) and that overgeneralization from lab experiments in many areas of behavioral science can (and routinely does) lead practitioners and policymakers to deploy suboptimal and even dangerous real-world interventions (Brewin, 2022; de Leeuw, Motz, Fyfe, Carvalho, & Goldstone, 2022; Grubbs, 2022; Wiernik, Raghavan, Allan, & Denison, 2022). Therefore, social scientists should precisely identify the most effective intervention under each arising set of circumstances.

The integrative approach naturally emphasizes contingencies and enables practitioners to distinguish between the *most general* result and the result that is *most useful in practice*. For example, in Figure 2B, the experiments depicted with a gray point correspond to the most general claim, occupying the largest region in the design space. However, this view ignores *relevance*, defined as points that represent the “target” conditions or the particular real-world context to which the practitioner hopes to generalize the results (Berkman & Wilson, 2021; Brunswik, 1955), as shown in Figure 2C. By concretely emphasizing these theoretical contingencies, the integrative approach supports “use-inspired” research (Stokes, 1997; Watts, 2017).

## 4. Existing steps toward integrative experiments

Integrative experiment design is not yet an established framework. However, some recent experimental work has begun to move in the direction we endorse – for example, by explicitly constructing a design space, sampling conditions more broadly and densely than the one-at-a-time approach would have, and constructing new kinds of theories that reflect the complexity of human behavior. In this section, we describe three examples of such experiments in the domains of (1) moral judgments, (2) risky choices, and (3) subliminal priming effects. Note that these examples are not an exhaustive accounting of relevant work, nor fully fleshed out exemplars of the integrative framework. Rather, we find them to be helpful illustrations of work that is closely adjacent to what we describe and evidence that the approach is realizable and can yield useful insights.

### 4.1. Factors influencing moral judgments

Inspired by the trolley problem, the seminal “Moral Machine” experiment used crowdsourcing to study human perspectives on moral decisions made by autonomous vehicles (Awad et al., 2018, 2020). The experiment was supported by an algorithm that sampled a nine-dimensional space of over 9 million distinct moral dilemmas. In the first 18 months after deployment, the

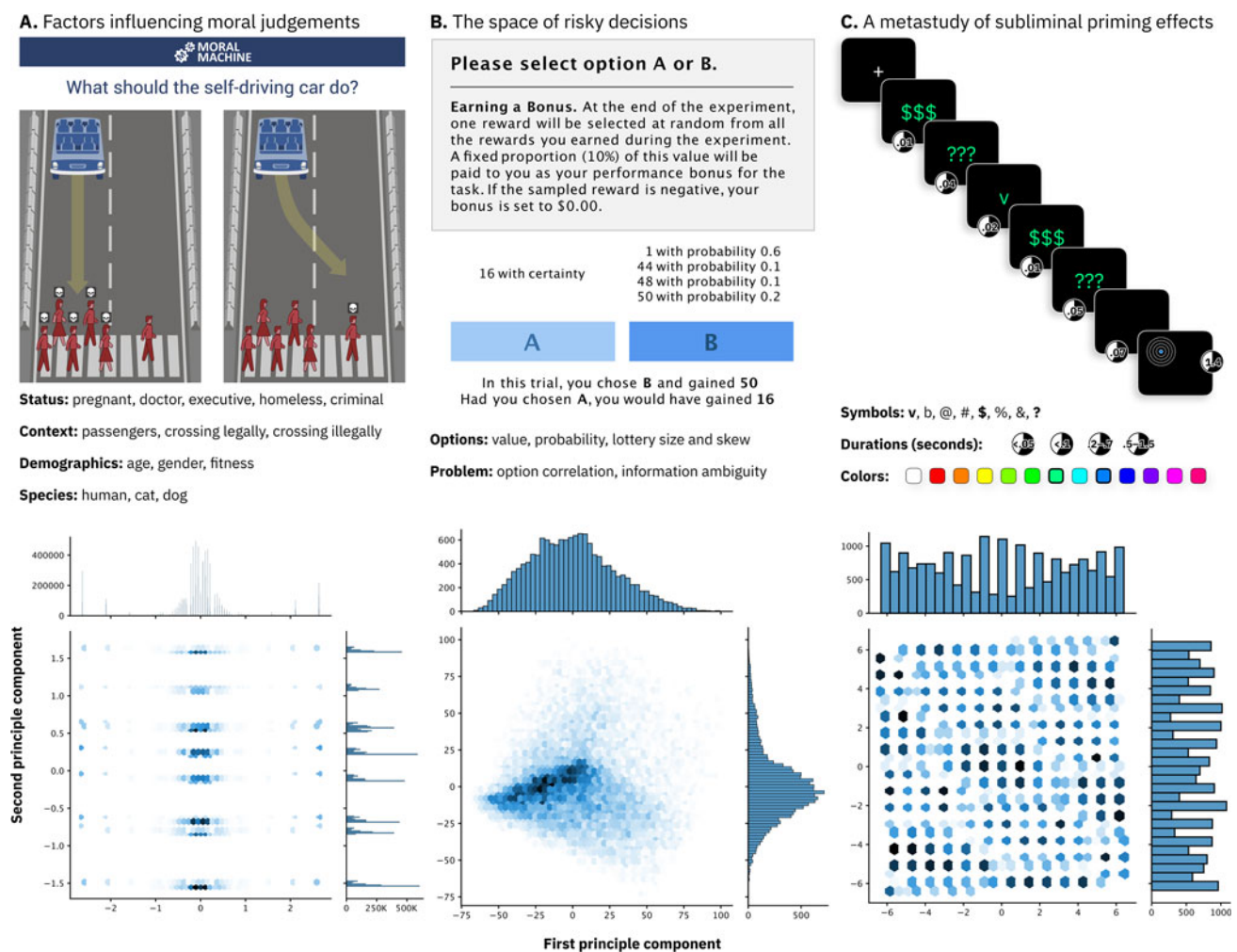


researchers collected more than 40 million decisions in 10 languages from over 4 million unique participants in 233 countries and territories (Fig. 3A).

The study offers numerous findings that were neither obvious nor deducible from prior research or traditional experimental designs. For example, they show that once a moral dilemma is made sufficiently complex, few people will hold to the principle of treating all lives equally. Instead, they appear to treat demographic groups quite differently – for example, a willingness to sacrifice the elderly in service of the young, and a preference for sparing the wealthy over the poor at about the same level as the preference for preserving people following the law over those breaking it (Awad et al., 2018). A second surprising finding by Awad et al. (2018) was that the differences between omission and commission (a staple of discussions of Western moral philosophy) ranks surprisingly low relative to other variables affecting judgments of morality and that this ethical preference for inaction is primarily concentrated in Western cultures (e.g., North America and many European countries of Protestant, Catholic, and Orthodox Christian cultural groups). Indeed, the observation that clustering between countries is not just based on one or two

ethical dimensions, but on a full profile of the multiplicity of ethical dimensions is something that would have been impossible to detect using studies that lacked the breadth of experimental conditions sampled in this study.

Moreover, such an approach to experimentation yields datasets that are more useful to other researchers as they evaluate their hypotheses, develop new theories, and address long-standing concerns such as which variables matter most to producing a behavior and what their relative contributions might be. For instance, Agrawal and colleagues used the dataset generated by the Moral Machine experiment to build a model with a black-box machine-learning method (specifically, an artificial neural network) for predicting people’s decisions (Agrawal, Peterson, & Griffiths, 2020). This predictive model was used to critique a traditional cognitive model and identify potentially causal variables influencing people’s decisions. The cognitive model was then evaluated in a new round of experiments that tested its predictions about the consequences of manipulating the causal variables. This approach of “scientific regret minimization” combined machine learning with rational choice models to jointly maximize the theoretical model’s predictive accuracy and interpretability in the context



**Figure 3.** Examples of integrative experiments. The top row illustrates the experimental tasks used in the Moral Machine, decisions under risk, and subliminal priming effects experiments, respectively, followed by the parameters varied across each experiment (bottom row). Each experiment instance (i.e., a scenario in the Moral Machine experiment, a pair of gambles in the risky-choice experiment, and a selection of facet values in the subliminal priming effects experiment) can be described by a vector of parameter values. Reducing the resulting space to two dimensions (2D) visualizes coverage by different experiments. This 2D embedding results from applying principal component analysis (PCA) to the parameters of these experimental conditions.

of moral judgments. It also yielded a more-complex theory than psychologists might be accustomed to: The final model had over 100 meaningful predictors, each of which could have been the subject of a distinct experiment and theoretical insight about human moral reasoning. By considering the influence of these variables in a single study by Awad et al. (2018), the researchers could ask what contribution each made to explaining the results. Investigation at this scale becomes possible when machine-learning methods augment the efforts of human theorists (Agrawal et al., 2020).

#### 4.2. The space of risky decisions

The choice prediction competitions studied human decisions under risk (i.e., where outcomes are uncertain) by automating selection of more than 100 pairs of gambles from a 12-dimensional space with an algorithm (Erev, Ert, Plonsky, Cohen, & Cohen, 2017; Plonsky et al., 2019). Recent work scaled this approach by taking advantage of the larger sample sizes made possible by virtual labs, collecting human decisions for over 10,000 pairs of gambles (Bourgin, Peterson, Reichman, Russell, & Griffiths, 2019; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021).

By sampling the space of possible experiments (in this case, gambles) much more densely (Fig. 3B), Peterson et al. (2021) found that two of the classic phenomena of risky choice – loss aversion and overweighting of small probabilities – did not manifest uniformly across the entire space of possible gambles. These two phenomena originally prompted the development of prospect theory (Kahneman & Tversky, 1979), representing significant deviations from the predictions of classic expected utility theory. By identifying regions of the space of possible gambles where loss aversion and overweighting of small probabilities occur, Kahneman and Tversky showed that expected utility theory does not capture some aspects of human decision making. However, in analyzing predictive performance across the entire space of gambles, Peterson et al. found that prospect theory was outperformed by a model in which the degree of loss aversion and overweighting of small probabilities varied smoothly over the space.

The work of Peterson et al. (2021) illustrates how the content of theories might be expected to change with a shift to the integrative approach. Prospect theory makes a simple assertion about human decision making: People exhibit loss aversion and overweight small probabilities. Densely sampling a larger region of the design space yields a more nuanced theory: While the functional form of prospect theory is well suited for characterizing human decisions, the extent to which people show loss aversion and overweight small probabilities depends on the context of the choice problem. That dependency is complicated. Even so, Peterson et al. identified several relevant variables such as the variability of the outcomes of the underlying gambles and whether the gamble was entirely in the domain of losses. Machine-learning methods were useful in developing this theory, initially to optimize the parameters of the functions assumed by prospect theory and other classic theories of decision making so as to ensure evaluation of the best possible instances of those theories, and then to demonstrate that these models did not capture variation in people's choices that could be predicted by more-complex models.

#### 4.3. A metastudy of subliminal priming effects

A recent cognitive psychology paper described an experiment in which a subliminal cue influences how participants balance

speed and accuracy in a response-time task (Reuss, Kiesel, & Kunde, 2015). In particular, participants were instructed to rapidly select a target according to a cue that signaled whether to prioritize response accuracy over speed, or vice versa. Reuss et al. reported typical speed–accuracy tradeoffs: When cued to prioritize speed, participants were faster and gave less accurate responses, whereas when cued to prioritize accuracy, participants were slower and more accurate. Crucially, this relationship was also found with cues that were rendered undetectable via a *mask*, an image presented directly before or after the cue that can suppress conscious perception of it.

The study design of the original experiment included several nuisance variables (e.g., the color of the cue), the values of which were not thought to affect the finding of subliminal effects. If the claimed effects were general, it would appear for all plausible values of the nuisance variables, whereas its appearance in some (contiguous) ranges of values but not in others would indicate contingency. And if spurious, the effect would appear only for the original values, if at all.

Baribault et al. (2018) took a “radical randomization” approach (also called a “metastudy” approach) in examining the generalizability and robustness of the original finding by randomizing 16 independent variables that could moderate the subliminal priming effect (Fig. 3C). By sampling nearly 5,000 “microexperiments” from the 16-dimensional design space, Baribault et al. revealed that masked cues had an effect on participant behavior only in the subregion of the design space where the cue is consciously visible, thus providing much stronger evidence about the lack of the subliminal priming effect than any single traditional experiment evaluating this effect could have. For a recent, thorough discussion of the metastudy approach and its advantages, along with a demonstration using the risky-choice framing effect, see DeKay, Rubinchik, Li, and De Boeck (2022).

### 5. Critiques and concerns

We have argued that adopting what we have called “integrative designs” in experimental social and behavioral science will lead to more-consistent, more-cumulative, and more-useful science. As should be clear from our discussion, however, our proposal is preliminary and therefore subject to several questions and concerns. Here we outline some of the critiques we have encountered and offer our responses.

#### 5.1. Isn't the critique of the one-at-a-time approach unfair?

One possible response is that our critique of the one-at-a-time approach is unduly critical and does not recognize its proper role in the future of social and behavioral sciences. To be clear, we are neither arguing that scientists should discard the “one-at-a-time” paradigm entirely nor denigrating studies (including our own!) that have employed it. The approach has generated a substantial amount of valuable work and continues to be useful for understanding individual causal effects, shaping theoretical models, and guiding policy. For example, it can be a sufficient and effective means to provide evidence for the existence of a phenomenon (but not the conditions under which it exists), as in field experiments that show that job applicants with characteristically “Black” names are less likely to be interviewed than those with “White” names, revealing the presence of structural racism and informing public debates about discrimination (Bertrand & Mullainathan, 2004). Moreover, one-at-a-time

experimentation can precede the integrative approach when exploring a new topic and identifying the variables that make up the design space.

Rather, our point is that the one-at-a-time approach cannot do all the work that is being asked of it, in large part because theories in the social and behavioral sciences cannot do all the work that is being asked of them. Once we recognize the inherent imprecision and ambiguity of social and behavioral theories, the lack of commensurability across independently designed and executed experiments is revealed as inevitable. Similarly, the solution we describe here can be understood simply as baking commensurability into the design process, by explicitly recognizing potential dimensions of variability and mapping experiments such that they can be compared with one another. In this way, the integrative approach can complement one-at-a-time experiments by incorporating them within design spaces (analogous to how articles already contextualize their contribution in terms of the prior literature), through which the research field might quickly recognize creative and pathbreaking contributions from one-at-a-time research.

### 5.2. Can't we solve the problem with meta-analysis?

As discussed earlier, meta-analyses offer the attractive proposition that accumulation of knowledge can be achieved through a procedure that compares and combines results across experiments. But the integrative approach is different in at least three important ways.

First, meta-analyses – as well as systematic reviews and integrative conceptual reviews – are by nature *post hoc* mechanisms for performing integration: The synthesis and integration steps occur after the data are collected and the results are published. Therefore, it can take years of waiting for studies to accumulate “naturally” before one can attempt to “put them together” via meta-analyses (if at all, as the vast majority of published effects are never meta-analyzed). More importantly, because commensurability is not a first-order consideration of one-at-a-time studies, attempts to synthesize collections of such studies after the fact are intrinsically challenging. The integrative approach is distinct in that it treats commensurability as a first-order consideration that is baked into the research design at the outset (i.e., *ex ante*). As we have argued, the main benefit of *ex ante* over *ex post* integration is that the explicit focus on commensurability greatly eases the difficulty of comparing different studies and hence integrating their findings (whether similar or different). In this respect, our approach can be viewed as a “planned meta-analysis” that is explicitly designed to sample conditions more broadly, minimize sampling bias, and efficiently reveal how effects vary across conditions. Although it may take more time and effort (and thus money) to run an integrative experiment than a single traditional experiment, when considering the accumulated effort of all the original research, this effort is much less than that of typical meta-analyses (see sect. 5.6 for a discussion about costs).

Second, although a meta-analysis typically aims to estimate the size of an effect by aggregating (e.g., averaging) over design variations across experiments, our emphasis is on trying to map the variation in an effect across an entire design space. While some meta-analyses with sufficient data attempt to determine the heterogeneity of the effect of interest, these efforts are typically hindered by the absence of systematic data on the variations in design choices (as well as in methods).

Third, publication bias induced by selective reporting of conditions and results – known as the file drawer problem (Carter,

Schönbrodt, Gervais, & Hilgard, 2019; Rosenthal, 1979) – can lead to biased effect-size estimates in meta-analyses. While there are methods for identifying and correcting such biases, one cannot be sure of their effectiveness in any particular case because of their sensitivity to untestable assumptions (Carter et al., 2019; Cooper, Hedges, & Valentine, 2019). Another advantage of the integrative approach is that it is largely immune to such problems because all sampled experiments are treated as informative, regardless of the novelty or surprise value of the individual findings, thereby greatly reducing the potential for bias.

### 5.3. How do integrative experiments differ from other recent innovations in psychology?

There have been several efforts to innovate on traditional experiments in the behavioral and social sciences. One key innovation is collaboration by multiple research labs to conduct systematic replications or to run larger-scale experiments than had previously been possible. For instance, the Many Labs initiative coordinated numerous research labs to conduct a series of replications of significant psychological results (Ebersole et al., 2016; Klein et al., 2014, 2018). This effort has itself been replicated in enterprises such as the ManyBabies Consortium (ManyBabies Consortium, 2020), ManyClasses (Fyfe et al., 2021), and ManyPrimates (Many Primates et al., 2019), which pursue the same goal with more-specialized populations, and in the DARPA SCORE program, which did so over a representative sample of experimental research in the behavioral and social sciences (Witkop, n.d.).<sup>13</sup> The Psychological Science Accelerator brings together multiple labs with a different goal: To evaluate key findings in a broader range of participant populations and at a global scale (Moshontz et al., 2018). Then, there is the Crowdsourcing Hypothesis Tests collaboration, which assigned 15 research teams to each design a study targeting the same hypothesis, varying in methods (Landy et al., 2020). Moreover, there is a recent trend in behavioral science to run “megastudies,” in which researchers test a large number of treatments in a single study in order to increase the pace and comparability of experimental results (Milkman et al., 2021, 2022; Voelkel et al., 2022).

All of these efforts are laudable and represent substantial methodological advances that we view as complements to, not substitutes for, integrative designs. What is core to the integrative approach is the explicit construction of, sampling from, and building theories upon a design space of experiments. Each ongoing innovation can contribute to the design of integrative experiments in its own way. For example, large-scale collaborative networks such as Many Labs can run integrative experiments together by assigning points in the design space to participating labs. Or in the megastudy research design, the interventions selected by researchers can be explicitly mapped into design spaces and then analyzed in a way that aims to reveal contingencies and generate metatheories of the sort discussed in section 3.3.

### 5.4. What about unknown unknowns?

There will always be systematic nontrivial variables that should be represented in the design space but are missing – these are the unknown unknowns. We believe our responses to this challenge are worth expanding upon.

First, we acknowledge the challenge inherent in the first step of integrative experiment design: Constructing the design space. This construction requires identifying the subset of variables to include



from an infinite set of possible variables that could define the design space of experiments within a domain. To illustrate such a process, we discussed the example domain of group synergy (see sect. 3.1). But, of course, we think that the field is wide open, with many options to explore; that the methodological details will depend on the domain of interest; and that best practices will emerge with experience.

Second, although we do not yet know which of the many potentially relevant dimensions should be selected to represent the space, and there are no guarantees that all (or even most) of the selected dimensions will play a role in determining the outcome, the integrative approach can shed light on both issues. On the one hand, experiments that map to the same point in the design space but yield different results indicate that some important dimension is missing from the representation of the space. On the other, experiments that systematically vary in the design space but yield similar results could indicate that the dimensions where they differ are irrelevant to the effect of interest and should be collapsed.

### 5.5. *This sounds great in principle but it is impossible to do in practice*

Even with an efficient sampling scheme, integrative designs are likely to require a much larger number of experiments than is typical in the one-at-a-time paradigm; therefore, practical implementation is a real concern. However, given recent innovations in virtual lab environments, participant sourcing, mass collaboration mechanisms, and machine-learning methods, the approach is now feasible to some.

#### 5.5.1. *Virtual lab environments*

Software packages such as jsPsych (de Leeuw, 2015) nodeGame (Baliotti, 2017), Dallinger (<https://dallinger.readthedocs.io/>), Pushkin (Hartshorne, de Leeuw, Goodman, Jennings, & O'Donnell, 2019), Hemlock (Bowen, n.d.), and Empirica (Almaatouq et al., 2021b) support development of integrative experiments that can systematically cover an experimental design's parameter space with automatically executed conditions. Even with these promising tools, for which development is ongoing, we still believe that one of the most promising, cost-effective ways to accelerate and improve progress in social science is to increase investment in automation (Yarkoni et al., 2019).

#### 5.5.2. *Recruiting participants*

Another logistical challenge to integrative designs is that adequately sampling the space of experiments will typically require a large participant pool from which the experimenter can draw, often repeatedly. As it stands, the most common means of recruiting participants online involves crowdsourcing platforms (Horton, Rand, & Zeckhauser, 2011; Mason & Suri, 2012). The large-scale risky-choice dataset described above, for example, used this approach to collect its 10,000 pairs of gambles (Bourgin et al., 2019). However, popular crowdsourcing platforms such as Amazon Mechanical Turk (Litman, Robinson, & Abberbock, 2017) were designed for basic labeling tasks, which can be performed by a single person and require low levels of effort. And the crowdworkers performing the tasks may have widely varying levels of commitment and produce work of varying quality (Goodman, Cryder, & Cheema, 2013). Researchers are prevented by Amazon's terms of use from knowing whether crowdworkers have participated in similar experiments in the

past, possibly as professional study participants (Chandler, Mueller, & Paolacci, 2014). To accommodate behavioral research's special requirements, Prolific and other services (Palan & Schitter, 2018) have made changes to the crowdsourcing model, such as by giving researchers greater control over how participants are sampled and over the quality of their work.

Larger, more diverse volunteer populations are also possible to recruit, as the Moral Machine experiment exemplifies. In the first 18 months after deployment, that team gathered more than 40 million moral judgments from over 4 million unique participants in 233 countries and territories (Awad, Dsouza, Bonnefon, Shariff, & Rahwan, 2020). Recruiting such large sample sizes from volunteers is appealing; however, success with such recruitment requires participant-reward strategies like gamification or personalized feedback (Hartshorne et al., 2019; Li, Germine, Mehr, Srinivasan, & Hartshorne, 2022). Thus, it has been hard to generalize the model to other important research questions and experiments, particularly when taking part in the experiment does not appear to be fun or interesting. Moreover, such large-scale data collection using viral platforms such as the Moral Machine may require some flexibility from Institutional Review Boards (IRBs), as they resemble software products that are open to consumers more than they do closed experiments that recruit from well-organized, intentional participant pools. In the Moral Machine experiment, for example, the MIT IRB approved pushing the consent to an "opt-out" option at the end, rather than obtaining consent prior to participation in the experiment, as the latter would have significantly increased participant attrition (Awad et al., 2018).

#### 5.5.3. *Mass collaboration*

Obtaining a sufficiently large sample may require leveraging emerging forms of organizing research in the behavioral and social sciences, such as distributed collaborative networks of laboratories (Moshontz et al., 2018). As we discussed earlier, in principle, large-scale collaborative networks can cooperatively run integrative experiments by assigning points in the design space to participating labs.

#### 5.5.4. *Machine learning*

The physical and life sciences have benefited greatly from machine learning. Astrophysicists use image-classification systems to interpret the massive amounts of data recorded by their telescopes (Shallue & Vanderburg, 2018). Life scientists use statistical methods to reconstruct phylogeny from DNA sequences and use neural networks to predict the folded structure of proteins (Jumper et al., 2021). Experiments in the social and behavioral sciences, in contrast, have had relatively few new methodological breakthroughs related to these technologies. While social and behavioral scientists in general have embraced "big data" and machine learning, their focus to date has largely been on nonexperimental data.<sup>14</sup> In contrast, the current scale of experiments in the experimental social and behavioral sciences does not typically produce data at the volumes necessary for machine-learning models to yield substantial benefits over traditional methods.

Integrative experiments offer several new opportunities for machine-learning methods to be used to facilitate social and behavioral science. First, by producing larger datasets – either within a single experiment or across multiple integrated experiments in the same design space – the approach makes it possible to use a wider range of machine-learning methods, particularly ones less constrained by existing theories. This advantage is



illustrated by the work of Peterson et al. (2021), whose neural network models were trained on human choice data to explore the implications of different theoretical assumptions for predicting decisions. Second, these methods can play a valuable role in helping scientists make sense of the many factors that potentially influence behavior in these larger datasets, as in Agrawal et al.'s (2020) analysis of the Moral Machine data. Finally, machine-learning techniques are a key part of designing experiments that efficiently explore large design spaces, as they are used to define surrogate models that are the basis for active sampling methods.

### 5.6. *Even if such experiments are possible, costs will be prohibitive*

It is true that integrative experiments are more expensive to run than individual one-at-a-time experiments, which may partly explain why the former have not yet become more popular. However, this comparison is misleading because it ignores the cost of human capital in generating scientific insight. Assume that a typical experimental paper in the social and behavioral sciences reflects on the order of \$100,000 of labor costs in the form of graduate students or postdocs designing and running the experiment, analyzing the data, and writing up the results. Under the one-at-a-time approach, such a paper typically contains just one or at most a handful of experiments. The next paper builds upon the previous results and the process repeats. With hundreds of articles published over a few decades, the cumulative cost of a research program that explores roughly 100 points in the implicit design space easily reaches tens of millions of dollars.

Of those tens of millions of dollars, a tiny fraction – on the order of \$1,000 per paper, or \$100,000 per research program (<1%) – is spent on data collection. If instead researchers conducted a single-integrative experiment that covered the entire design space, they could collect all the data produced by the entire research program and then some. Even if this effort explored the design space significantly less efficiently than the traditional research program, requiring 10 times more data, data collection would cost about \$1,000,000 (<10%). This is a big financial commitment, but the labor costs for interpreting these data do not scale with the amount of data. So, even if researchers needed to commit 10 times as much labor as for a typical research paper, they would have discovered everything an entire multidecade research program would uncover in a single study costing only \$2,000,000.

The cost–benefit ratio of integrative experiments is hence at least an order of magnitude better than that of one-at-a-time experiments.<sup>15</sup> Pinching pennies on data collection results in losing dollars (and time and effort) in labor. If anything, when considered in aggregate, the efficiency gains of the integrative approach will be substantially greater than this back of the envelope calculation suggests. As an institution, the social and behavioral sciences have spent tens of billions of dollars during the past half-century.<sup>16</sup> With integrative designs, a larger up-front investment can save decades of unfruitful investigation and instead realize grounded, systematic results.

### 5.7. *Does this mean that small labs can't participate?*

Although the high up-front costs of designing and running an integrative experiment may seem to exclude small labs as well

as Principal investigators (PIs) from low-resource institutions, we anticipate that the integrative approach will actually broaden the range of people involved in behavioral research. The key insight here is that the methods and infrastructure needed to run integrative experiments are inherently shareable. Thus, while the development costs are indeed high, once the infrastructure has been built, the marginal costs of using it are low – potentially even lower than running a single, one-at-a-time experiment. As long as funding for the necessary technical infrastructure is tied to a requirement for sustaining collaborative research (as discussed in previous sections), it will create opportunities for a wider range of scientists to be involved in integrative projects and for researchers at smaller or undergraduate-focused institutions to participate in ambitious research efforts.

Moreover, research efforts in other fields illustrate how labs of different sizes can make different kinds of contributions. In biology and physics, some groups of scientists form consortia that work together to define a large-scale research agenda and seek the necessary funding (as described earlier, several thriving experimental consortia in the behavioral sciences illustrate this possibility). Other groups develop theory by digging deeper into the data produced by these large-scale efforts to make discoveries they may not have imagined when the data were first collected; some scientists focus on answering questions that do not require large-scale studies, such as the properties of specific organisms or materials that can be easily studied in a small lab; still other researchers conduct exploratory work to identify the variables or theoretical principles that may be considered in future large-scale studies. We envision a similar ecosystem for the future of the behavioral sciences.

### 5.8. *Shouldn't the replication crisis be resolved first?*

The replication crisis in the behavioral sciences has led to much reflection about research methods and substantial efforts to conduct more-applicable research (Freese & Peterson, 2017). We view our proposal as being consistent with these goals, but with a different emphasis than replication. To some extent, this difference is complementary to replication and can be pursued in parallel with it, but may suggest a different allocation of resources than a “replication first” approach.

Discussing the complementary role first, integrative experiments naturally support replicable science. Because choices about nuisance variables are rarely documented systematically in the one-at-a-time paradigm, it is not generally possible to establish how similar or different two experiments are. This observation may account for some recently documented replication failures (Camerer et al., 2018; Levinthal & Rosenkopf, 2021). While the replication debate has focused on shoddy research practices (e.g., *p*-hacking) and bad incentives (e.g., journals rewarding “positive, novel, and exciting” results), another possible cause of nonreplication is that the replicating experiment is in fact sufficiently dissimilar to the original (usually as a result of different choices of nuisance parameters) that one should not expect the result to replicate (Muthukrishna & Henrich, 2019; Yarkoni, 2022). In other words, without operating within a space that makes experiments commensurate, failures to replicate previous findings are never conclusive, because doubt remains as to whether one of the many possible moderator variables explains the lack of replication (Cesario, 2014). Regardless of whether an experimental finding's fragility to (supposedly) theoretically irrelevant parameters should be considered a legitimate defense of the

finding, the difficulty of resolving such arguments further illustrates the need for a more explicit articulation of theoretical scope conditions.

The integrative approach, accepting that treatment effects vary across conditions, would also recommend that directing massive resources to replicating existing effects may not be the best way to help our fields advance. Given that those historical effects were discovered under the one-at-a-time approach, they evaluate only specific points in the design space. Consistent with the argument above, rather than trying to perfectly reproduce those points in the design space (via “direct” replications), a better use of resources would be to sample the design space more extensively and use continuous measures to compare different studies (Gelman, 2018). In this way, researchers can not only discover whether historical effects replicate, but also draw stronger conclusions about whether (and to what extent) they generalize.

### 5.9. This proposal is incompatible with incentives in the social and behavioral sciences

Science does not occur in a vacuum. Scientists are constantly evaluated by their peers as they submit papers for publication, seek funding, apply for jobs, and pursue promotions. For the integrative approach to become widespread, it must be compatible with the incentives of individual behavioral scientists, including early career researchers. Given the current priority that hiring, tenure & promotion, and awards committees in the social and behavioral sciences place on identifiable individual contributions (e.g., lead authorship of scholarly works, perceived “ownership” of distinct programs of research, leadership positions, etc.), a key pragmatic concern is that the large-scale collaborative nature of integrative research designs might make them less rewarding than the one-at-a-time paradigm for anyone other than the project leaders.

Although a shift to large-scale, collaborative science does indeed present an adoption challenge, it is encouraging to note that even more dramatic shifts have taken place in other fields. In physics, for example, some of the most important results in recent decades – the discovery of the Higgs Boson (Aad et al., 2012), gravitational waves (Abbott et al., 2016), and so on – have been obtained via collaborations of thousands of researchers.<sup>17</sup> To ensure that junior team members are rewarded for their contributions, many collaborations maintain “speaker lists” that prominently feature early career researchers, offering them a chance to appear as the face of the collaboration. When these researchers apply for jobs or are considered for promotion, the leader of the collaboration writes a letter of recommendation that describes the scientists’ role in the collaboration and why their work is significant. A description of such roles can also be included directly in manuscripts through the Contributor Roles Taxonomy (Allen, Scott, Brand, Hlava, & Altman, 2014), a high-level taxonomy with 14 roles that describe typical contributions to scholarly output; the taxonomy has been adopted as an American National Standards Institute (ANSI)/National Information Standards Organization (NISO) standard and is beginning to see uptake (National Information Standards Organization, 2022). Researchers who participate substantially in creating the infrastructure used by a collaborative effort can receive “builder” status, appearing as coauthors on subsequent publications that use that infrastructure. Many collaborations also have mentoring plans designed to support early career researchers. Together, these mechanisms are intended to make participation in large collaborations attractive to a wide range of researchers at various career

stages. While acknowledging that physics differs in many ways from the social and behavioral sciences, we nonetheless believe that the model of large collaborative research efforts can take root in the latter. Indeed, we have already noted the existence of several large collaborations in the behavioral sciences that appear to have been successful in attracting participation from small labs and early career researchers.

## 6. Conclusion

The widespread approach of designing experiments one-at-a-time – under different conditions with different participant pools, and with nonstandardized methods and reporting – is problematic because it is at best an inefficient way to accumulate knowledge, and at worst it fails to produce consistent, cumulative knowledge. The problem clearly will not be solved by increasing sample sizes, focusing on effect sizes rather than statistical significance, or replicating findings with preregistered designs. We instead need a fundamental shift in how to think about theory construction and testing.

We describe one possible approach, one that promotes commensurability and continuous integration of knowledge by design. In this “integrative” approach, experiments would not just evaluate a few hypotheses but would explore and integrate over a wide range of conditions that deserve explanation by all pertinent theories. Although this kind of experiment may strike many as atheoretical, we believe the one-at-a-time approach owes its dominance not to any particular virtues of theory construction and evaluation but rather to the historical emergence of experimental methods under a particular set of physical and logistical constraints. Over time, generations of researchers have internalized these features to such an extent that they are thought to be inseparable from sound scientific practice. Therefore, the key to realizing our proposed type of reform – and to making it productive and useful – is not only technical, but also cultural and institutional.

**Acknowledgments.** We owe an important debt to Saul Perlmutter, Serguei Saavedra, Matthew J. Salganik, Gary King, Todd Gureckis, Alex “Sandy” Pentland, Thomas W. Malone, David G. Rand, Iyad Rahwan, Ray E. Reagans, and the members of the MIT Behavioral Lab and the UPenn Computational Social Science Lab for valuable discussions and comments. This article also benefited from conversations with dozens of people at two workshops: (1) “Scaling Cognitive Science” at Princeton University in December 2019, and (2) “Scaling up Experimental Social, Behavioral, and Economic Science” at the University of Pennsylvania in January 2020.

**Financial support.** This work was supported in part by the Alfred P. Sloan Foundation (2020-13924) and the NOMIS Foundation.

**Competing interest.** None.

## Notes

1. Although we restrict the focus of our discussion to lab experiments in the social and behavioral sciences, with which we are most familiar, we expect that our core arguments generalize well to other modes of inquiry and adjacent disciplines.
2. By analogy, we note that for almost as long as  $p$ -values have been used as a standard of evidence in the social and behavioral sciences, critics have argued that they are somewhere between insufficient and meaningless (Cohen, 1994; Dienes, 2008; Gelman & Carlin, 2017; Meehl, 1990a). Yet, in the absence of an equally formulaic alternative,  $p$ -value analysis remains pervasive (Benjamin et al., 2018).
3. Nor do recent proposals to improve the replicability and reproducibility of scientific results (Gelman & Loken, 2014; Ioannidis, 2005; Munafò et al., 2017;

Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011) address the problem. While these proposals are worthy, their focus is on individual results, not on how collections of results fit together.

4. We also note that in an alternative formulation of the design space, all variables (including what one would think of as experimental manipulations) are included as dimensions of the design space and the focal experimental manipulation is represented as a comparison across two or more points in the space. Some of the examples described in section 4 are more readily expressed in one formulation, whereas others are more readily expressed in the other. They are equivalent: It is possible to convert from one to the other without any loss of information.

5. To illustrate with another example, cultural psychologists such as Hofstede (2016), Inglehart and Welzel (2005), and Schwartz (2006) identified cultural dimensions along which groups differ, which then can be used to define distance measures between populations and to guide researchers in deciding where to target their data-collection efforts (Muthukrishna et al., 2020). Another example of this exercise is the extensive breakdown of the “auction design space” by Wurman, Wellman, and Walsh (2001), which captures the essential similarities and differences of many auction mechanisms in a format more descriptive and useful than simple taxonomies and serves as an organizational framework for classifying work within the field.

6. Active learning is also called “query learning” or sometimes “sequential optimal experimental design” in the statistics literature.

7. Active learning has recently become an important tool for optimizing experiments in other fields, such as machine-learning hyperparameters (Snoek, Larochelle, & Adams, 2012), materials and mechanical designs (Burger et al., 2020; Gongora et al., 2020; Lei et al., 2021), and chemical reaction screening (Eyke, Green, & Jensen, 2020, 2021; Shields et al., 2021) – just to mention a few.

8. For example, surrogate models can be probabilistic models (e.g., a Gaussian process) as well as nonprobabilistic (e.g., neural networks, tree-based methods), while sampling strategies can include uncertainty sampling, greedy sampling, and distance-based sampling.

9. Popular active learning libraries for experiments include Ax (Bakshy et al., 2018), BoTorch (Balandat et al., 2020), and GPflowOpt (Knudde, van der Herten, Dhaene, & Couckuyt, 2017).

10. See Settles (2011), Greenhill, Rana, Gupta, Vellanki, and Venkatesh (2020), and Ren et al. (2021) for surveys on active learning.

11. Given that the data from the integrative approach are generated independent of the current set of theories in the field, the resulting data are potentially informative not just about those theories, but about theories that are yet to be proposed. As a consequence, data generated by this integrative approach are intended to have greater longevity than data generated by “one-at-a-time” experiments.

12. Another explanation for the inability to make accurate predictions is that the majority of dimensions defining the design space are uninformative and need to be reconsidered.

13. For a more comprehensive list, see Uhlmann et al. (2019).

14. For example, the CHILDES dataset of child-directed speech (MacWhinney, 2014) has had a significant impact on studies of language development, and census data, macroeconomic data, and other large datasets (e.g., from social media and e-commerce platforms) are increasingly prevalent in political science, sociology, and economics.

15. This shift has already occurred in some areas. For example, the cognitive neuroscience field has been transformed in the past few decades by the availability of increasingly effective methods for brain imaging. Researchers now take for granted that data collection costs tens or hundreds of thousands of dollars and that the newly required equipment and other infrastructure for this kind of research costs millions of dollars – that is, they now budget more for data collection than for hiring staff. Unlocking the full potential of our envisioned integrative approach will require similarly new, imaginative ways of allocating resources and a willingness to spend money on generating more-definitive, reusable datasets (Griffiths, 2015).

16. The budget associated with the NSF Directorate for Social, Behavioral, and Economic Sciences alone is roughly 5 billion dollars over the past two decades and, by its 2022 estimate, accounts for “approximately 65 percent of the federal funding for basic research at academic institutions in the social, behavioral, and economic sciences” (National Science Foundation, 2022). Extending the time range to 50 years and accounting for sources of funding beyond the

US federal government, including all other governments, private foundations, corporations, and direct funding from universities, brings our estimate to tens of billions of dollars.

17. We thank Saul Perlmutter for sharing his perspective on how issues of incentives are addressed in physics, drawing on his experience in particle physics and cosmology.

## References

- Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., ... Zwalinski, L. (2012). Observation of a new particle in the search for the Standard Model Higgs Boson with the ATLAS detector at the LHC. *Physics Letters, Part B*, 716(1), 1–29.
- Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., ... LIGO Scientific Collaboration and Virgo Collaboration. (2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, 116(6), 061102.
- Aggarwal, I., & Woolley, A. W. (2018). Team creativity, cognition, and cognitive style diversity. *Management Science*, 65(4), 1586–1599. <https://doi.org/10.1287/mnsc.2017.3001>
- Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 117(16), 8825–8835.
- Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, 508(7496), 312–313.
- Allen, N. J., & Hecht, T. D. (2004). The “romance of teams”: Toward an understanding of its psychological underpinnings and implications. *Journal of Occupational and Organizational Psychology*, 77(4), 439–461.
- Allport, F. H. (1924). The group fallacy in relation to social science. *The American Journal of Sociology*, 29(6), 688–706.
- Almaatouq, A. (2019). Towards stable principles of collective intelligence under an environment-dependent framework. Massachusetts Institute of Technology. <https://dspace.mit.edu/handle/1721.1/123223?show=full?show=full>
- Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021a). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences of the United States of America*, 118(36), e2101062118. <https://doi.org/10.1073/pnas.2101062118>
- Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021b). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, 53, 2158–2171. <https://doi.org/10.3758/s13428-020-01535-9>
- Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 117(21), 11379–11386.
- Almaatouq, A., Rahimian, M. A., Burton, J. W., & Alhajri, A. (2022). The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Scientific Reports*, 12(1), 16546.
- Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., Duguid, S. J., ... Watzek, J. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLoS ONE*, 14(10), e0223675.
- Arrow, H., McGrath, J. E., & Berdahl, J. L. (2000). *Small groups as complex systems: Formation, coordination, development, and adaptation*. Sage.
- Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs (Oxford statistical science series, 8)* (1st ed.). Clarendon Press.
- Aumann, R. J., & Hart, S. (1992). *Handbook of game theory with economic applications*. Elsevier.
- Auspurg, K., & Hinz, T. (2014). *Factorial survey experiments*. Sage.
- Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, 63(3), 48–55.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The Moral Machine experiment. *Nature*, 563(7729), 59–64.
- Bakshy, E., Dworkin, L., Karrer, B., Kashin, K., Letham, B., Murthy, A., & Singh, S. (2018). AE: A domain-agnostic platform for adaptive experimentation. *Workshop on System for ML*. <http://learningsys.org/nips18/assets/papers/87CameraReadySubmissionAE%20-%20NeurIPS%202018.pdf>
- Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., & Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)* (pp. 21524–21538). Curran Associates Inc.
- Balietti, S. (2017). NodeGame: Real-time, synchronous, online experiments in the browser. *Behavior Research Methods*, 49(5), 1696–1715.
- Balietti, S., Klein, B., & Riedl, C. (2021). Optimal design of experiments to identify latent behavioral types. *Experimental Economics*, 24, 772–799. <https://doi.org/10.1007/s10683-020-09680-w>
- Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2607–2612.



- Barron, B. (2003). When smart groups fail. *Journal of the Learning Sciences*, 12(3), 307–359.
- Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 114(26), E5070–E5076.
- Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *The Journal of Applied Psychology*, 92(3), 595–615.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10. <https://doi.org/10.1038/s41562-017-0189-z>
- Berkman, E. T., & Wilson, S. M. (2021). So useful as a good theory? The practicality crisis in (social) psychological theory. *Perspectives on Psychological Science*, 16(4), 864–874. <https://doi.org/10.1177/1745691620969650>
- Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, 94(4), 991–1013.
- Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019). Cognitive model priors for predicting human decisions. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 5133–5141). PMLR.
- Bowen, D. (n.d.). Hemlock. Retrieved April 22, 2022, from <https://dsbowen.github.io/hemlock>
- Brewin, C. R. (2022). Impact on the legal system of the generalizability crisis in psychology. *The Behavioral and Brain Sciences*, 45, e7.
- Breznan, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., ... Zóttak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, 119(44), e2203150119.
- Brunswik, E. (1947). Systematic and representative design of psychological experiments. In *Proceedings of the Berkeley symposium on mathematical statistics and probability* (pp. 143–202). University of California Press.
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217.
- Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., ... Cooper, A. I. (2020). A mobile robotic chemist. *Nature*, 583(7815), 237–241.
- Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Kiley Hamlin, J., Kline, M., ... Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, 61(4), 349–363. <https://doi.org/10.1037/cap0000216>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644.
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144.
- Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, 9(1), 40–48. <https://doi.org/10.1177/1745691613513470>
- Cesario, J. (2022). What can experimental studies of bias tell us about real-world group disparities?. *Behavioral and Brain Sciences*, 45, E66. <https://doi.org/10.1017/S0140525X21000017>
- Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46(1), 112–130.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *The American Psychologist*, 49(12), 997.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- DeBrouwere, S., & Rosseel, Y. (2022). The conceptual, cunning and conclusive experiment in psychology. *Perspectives on Psychological Science*, 17(3), 852–862. <https://doi.org/10.1177/17456916211026947>
- DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating psychological science with metastudies: A demonstration using the risky-choice framing effect. *Perspectives on Psychological Science*, 17(6), 1704–1736. <https://doi.org/10.1177/17456916221079611>
- de Leeuw, J. R. (2015). JsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12.
- de Leeuw, J. R., Motz, B. A., Fyfe, E. R., Carvalho, P. F., & Goldstone, R. L. (2022). Generalizability, transferability, and the practice-to-practice gap [Review of *Generalizability, transferability, and the practice-to-practice gap*]. *The Behavioral and Brain Sciences*, 45, e11.
- Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, 7(3), 622–727.
- Devine, D. J., & Phillips, J. L. (2001). Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research*, 32(5), 507–532.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan.
- Dubova, M., Moskvichev, A., & Zollman, K. (2022). Against theory-motivated experimentation in science. *MetaArXiv*. June 24. <https://doi.org/10.31222/osf.io/yvs2u>
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82.
- Ellemers, N., & Rink, F. (2016). Diversity in work groups. *Current Opinion in Psychology*, 11, 49–53.
- Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS ONE*, 9(12), e115212.
- Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, 124(4), 369–409.
- Eyke, N. S., Green, W. H., & Jensen, K. F. (2020). Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering*, 5(10), 1963–1972.
- Eyke, N. S., Koscher, B. A., & Jensen, K. F. (2021). Toward machine learning-enhanced high-throughput experimentation. *Trends in Chemistry*, 3(2), 120–132.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4), 980–994.
- Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, 43, 147–165. <https://doi.org/10.1146/annurev-soc-060116-053450>
- Fyfe, E. R., de Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., ... Motz, B. A. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science*, 4(3), 25152459211027575.
- Gale, D., & Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1), 9–15.
- Gelman, A. (2018). Don't characterize replications as successes or failures [Review of *Don't characterize replications as successes or failures*]. *The Behavioral and Brain Sciences*, 41, e128.
- Gelman, A., & Carlin, J. (2017). Some natural solutions to the p-value communication problem – and why they won't work. *Journal of the American Statistical Association*, 112(519), 899–901.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis – a “garden of forking paths” – explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6), 460.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), 1–58.
- Gongora, A. E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K. G., ... Brown, K. A. (2020). A Bayesian experimental autonomous researcher for mechanical design. *Science Advances*, 6(15), eaaz1708.
- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples: Data collection in a flat world. *Journal of Behavioral Decision Making*, 26(3), 213–224.
- Greenhill, S., Rana, S., Gupta, S., Vellanki, P., & Venkatesh, S. (2020). Bayesian optimization for adaptive experimental design: A review. *IEEE Access*, 8, 13937–13948.
- Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, 135, 21–23.
- Grubbs, J. B. (2022). The cost of crisis in clinical psychological science [Review of *The cost of crisis in clinical psychological science*]. *The Behavioral and Brain Sciences*, 45, e18.
- Hackman, J. R. (1968). Effects of task characteristics on group products. *Journal of Experimental Social Psychology*, 4(2), 162–187.
- Harkins, S. G. (1987). Social loafing and social facilitation. *Journal of Experimental Social Psychology*, 23(1), 1–18.
- Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods*, 51(4), 1782–1803. <https://doi.org/10.3758/s13428-018-1155-z>
- Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, 33(2-3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560.
- Hill, G. W. (1982). Group versus individual performance: Are  $N + 1$  heads better than one? *Psychological Bulletin*, 91(3), 517–539.
- Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science (New York, N.Y.)*, 355(6324), 486–488.
- Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181–188.
- Hofstede, G. (2016). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations (2nd ed.). *Collegiate Aviation Review*, 34(2), 108–109. Retrieved from <https://www.proquest.com/scholarly-journals/cultures-consequences-comparing-values-behaviors/docview/1841323332/se-2>
- Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46), 16385–16389.






- Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3), 399–425.
- Husband, R. W. (1940). Cooperative versus solitary problem solution. *The Journal of Social Psychology*, 11(2), 405–409.
- Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes* (p. 277). Houghton Mifflin Company. <https://psycnet.apa.org/fulltext/1975-29417-000.pdf>
- Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., ... Coles, N. A. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*, 5(1), 159–169.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47(2), 263–291.
- Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706.
- Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What makes a strong team?: Using collective intelligence to predict team performance in league of legends. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing – CSCW '17* (pp. 2316–2329). New York, NY, USA.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, 45(3), 142–152.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490.
- Knudde, N., van der Herten, J., Dhaene, T., & Couckuyt, I. (2017). GPflowOpt: A Bayesian optimization library using TensorFlow. *arXiv [stat.ML]*. <http://arxiv.org/abs/1711.03845>
- Koyré, A. (1953). An experiment in measurement. *Proceedings of the American Philosophical Society*, 97(2), 222–237.
- Lakens, D., Uygun Tunç, D., & Necip Tunç, M. (2022). There is no generalizability crisis [Review of *There is no generalizability crisis*]. *The Behavioral and Brain Sciences*, 45, e25.
- Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5), 451–479.
- Larson, J. R. (2013). *In search of synergy in small group performance*. Psychology Press.
- Larson, S. D., & Martone, M. E. (2009). Ontologies for neuroscience: What are they and what are they good for? *Frontiers in Neuroscience*, 3(1), 60–67. <https://doi.org/10.3389/neuro.01.007.2009>
- Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, 88(2), 605–620.
- Lei, B., Kirk, T. Q., Bhattacharya, A., Pati, D., Qian, X., Arroyave, R., & Mallick, B. K. (2021). Bayesian optimization with adaptive surrogate models for automated experimental design. *NPJ Computational Materials*, 7(1), 1–12.
- LePine, J. A. (2003). Team adaptation and postchange performance: Effects of team composition in terms of members' cognitive ability and personality. *The Journal of Applied Psychology*, 88(1), 27–39.
- Letham, B., Karrer, B., Ottoni, G., & Bakshy, E. (2019). Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2), 495–519. <https://doi.org/10.1214/18-ba1110>
- Levinthal, D. A., & Rosenkopf, L. (2021). Commensurability and collective impact in strategic management research: When non-replicability is a feature, not a bug. Working-paper (unpublished preprint). <https://mackinstitute.wharton.upenn.edu/2020/commensurability-and-collective-impact-in-strategic-management-research/>
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives: A Journal of the American Economic Association*, 21(2), 153–174.
- Li, W., Germaine, L. T., Mehr, S. A., Srinivasan, M., & Hartshorne, J. (2022). Developmental psychologists should adopt citizen science to improve generalization and reproducibility. *Infant and Child Development*, e2348. <https://doi.org/10.1002/icd.2348>
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433–442.
- MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume II: The database* (3rd ed.). Psychology Press. <https://doi.org/10.4324/9781315805641>
- Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences of the United States of America*, 119(31), e2200300119.
- ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, 3(1), 24–52.
- Manzi, J. (2012). *Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society* (pp. 1–320). Basic Books.
- Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PLoS ONE*, 11(4), e0153048.
- Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). *Exploring limits to prediction in complex social systems*. In Proceedings of the 25th international conference on world wide web no. 978-1-4503-4143-1 (pp. 683–694). Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1–23.
- Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 109(3), 764–769.
- McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, 2(1), 3–19.
- McGrath, J. E. (1984). *Groups: Interaction and performance*. Prentice Hall.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115.
- Meehl, P. E. (1990a). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195–244.
- Meehl, P. E. (1990b). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108–141.
- Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences of the United States of America*, 119(1). <https://doi.org/10.1073/pnas.2107346118>
- Merton, R. K. (1968). On sociological theories of the middle range. *Social Theory and Social Structure*, 39–72.
- Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., ... Duckworth, A. L. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences of the United States of America*, 119(6). <https://doi.org/10.1073/pnas.2115126119>
- Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., ... Duckworth, A. L. (2021). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences of the United States of America*, 118(20), e2101165118.
- Mook, D. G. (1983). In defense of external invalidity. *The American Psychologist*, 38(4), 379–387.
- Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, 1(4), 501–515.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 21.
- Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, 31(6), 678–701.
- Muthukrishna, M., & Henrich, J. A. (2019). A problem in theory. *Nature Human Behaviour*, 3, 221–229. <https://doi.org/10.1038/s41562-018-0522-1>
- Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, 6(1), 58–73.
- National Information Standards Organization. (2022). ANSI/NISO Z39. 104-2022, CRediT, contributor roles taxonomy. [S. L.]. National Information Standards Organization. <https://www.niso.org/publications/z39104-2022-credit>
- National Science Foundation. (2022). NSF budget requests to congress and annual appropriations. National Science Foundation. <https://www.nsf.gov/about/budget/>
- Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports*, 11(1), 1980.
- Newell, A. (1973). *You can't play 20 questions with nature and win: Projective comments on the papers of this symposium*. <http://shelf2.library.cmu.edu/Tech/240474311.pdf>
- Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, 349(6251), aac4716.
- Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies – New edition*. Princeton University Press.
- Palan, S., & Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, 372(6547), 1209–1214.

- Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., ... Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv [cs.AI]. arXiv*. <http://arxiv.org/abs/1904.06866>
- Preckel, F., & Brunner, M. (2017). Nomological nets. *Encyclopedia of Personality and Individual Differences*, 1–4. [https://doi.org/10.1007/978-3-319-28099-8\\_1334-1](https://doi.org/10.1007/978-3-319-28099-8_1334-1)
- Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., ... Wang, X. (2021). A survey of deep active learning. *ACM Computing Surveys*, 54(9), 1–40.
- Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, 134, 57–62.
- Richard Hackman, J., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 8, pp. 45–99). Academic Press. [https://doi.org/10.1016/s0065-2601\(08\)60248-8](https://doi.org/10.1016/s0065-2601(08)60248-8)
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., ... Musen, M. A. (2006). National center for biomedical ontology: Advancing biomedicine through structured organization of scientific knowledge. *OMICS: A Journal of Integrative Biology*, 10(2), 185–198. <https://doi.org/10.1089/omi.2006.10.185>
- Schneid, M., Isidor, R., Li, C., & Kabst, R. (2015). The influence of cultural context on the relationship between gender diversity and team performance: A meta-analysis. *The International Journal of Human Resource Management*, 26(6), 733–756.
- Schulz-Hardt, S., & Mojzisch, A. (2012). How to achieve synergy in group decision making: Lessons to be learned from the hidden profile paradigm. *European Review of Social Psychology*, 23(1), 305–343.
- Schwartz, S. (2006). A theory of cultural value orientations: Explication and applications. *Comparative Sociology*, 5(2–3), 137–182.
- Settles, B. (2011). From theories to queries: Active learning in practice. In I. Guyon, G. Cawley, G. Dror, V. Lemaire, & A. Statnikov (Eds.), *Active learning and experimental design workshop in conjunction with AISTATS 2010* (Vol. 16, pp. 1–18). PMLR.
- Shallue, C. J., & Vanderburg, A. (2018). Identifying exoplanets with deep learning: A five-planet resonant chain around Kepler-80 and an eighth planet around Kepler-90. *AJS: American Journal of Sociology*, 155(2), 94.
- Shaw, M. E. (1963). *Scaling group tasks: A method for dimensional analysis*. <https://apps.dtic.mil/sti/pdfs/AD0415033.pdf>
- Shields, B. J., Stevens, J., Li, J., Parasram, M., Damani, F., Alvarado, J. I. M., ... Doyle, A. G. (2021). Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844), 89–96.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12(6), 1123–1128.
- Simonsohn, U., Simmons, J., & Nelson, L. D. (2022). Above averaging in literature reviews. *Nature Reviews Psychology*, 1(10), 551–552.
- Smucker, B., Krzywinski, M., & Altman, N. (2018). Optimal experimental design. *Nature Methods*, 15(8), 559–560.
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *arXiv [stat.ML]. arXiv*. <http://arxiv.org/abs/1206.2944>
- Steiner, I. D. (1972). *Group process and productivity*. Academic Press.
- Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management*, 32(1), 29–55.
- Stokes, D. E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.
- Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., ... Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions [Review of *No reason to expect large and consistent effects of nudge interventions*]. *Proceedings of the National Academy of Sciences of the United States of America*, 119(31), e2200732119.
- Tasca, G. A. (2021). Team cognition and reflective functioning: A review and search for synergy. *Group Dynamics: Theory, Research, and Practice*, 25(3), 258–270.
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4), 285–294.
- Turner, J. A., & Laird, A. R. (2012). The cognitive paradigm ontology: Design and application. *Neuroinformatics*, 10(1), 57–66.
- Turner, M. A., & Smaldino, P. E. (2022). Mechanistic modeling for the masses [Review of *Mechanistic modeling for the masses*]. *The Behavioral and Brain Sciences*, 45, e33.
- Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., ... Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 14(5), 711–733.
- Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23), 6454–6459.
- Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1), 8–37.
- Voelkel, J. G., Stagnaro, M. N., Chu, J., Pink, S. L., Mernyk, J. S., Redekopp, C., ... Willer, R. (2022). Megastudy identifying successful interventions to strengthen Americans' democratic attitudes. Preprint. <https://doi.org/10.31219/osf.io/y79u5>
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
- Watson, G. B. (1928). Do groups think more efficiently than individuals? *Journal of Abnormal and Social Psychology*, 23(3), 328.
- Watts, D. (2017). Response to Turco and Zuckerman's "Verstehen for sociology." *The American Journal of Sociology*, 122(4), 1292–1299.
- Watts, D. J. (2011). *Everything is obvious\*: Once you know the answer*. Crown Business.
- Watts, D. J. (2014). Common sense and sociological explanations. *The American Journal of Sociology*, 120(2), 313–351.
- Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, 1, 15.
- Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., Grubestic, A., ... Salganik, M. (2018). *Explanation, prediction, and causality: Three sides of the same coin?* <https://doi.org/10.31219/osf.io/u6vz5>
- Wiernik, B. M., Raghavan, M., Allan, T., & Denison, A. J. (2022). Generalizability challenges in applied psychological and organizational research and practice [Review of *Generalizability challenges in applied psychological and organizational research and practice*]. *The Behavioral and Brain Sciences*, 45, e38.
- Witkop, G. (n.d.). Systematizing confidence in open research and evidence (SCORE). DARPA. Retrieved June 22, 2022, from <https://www.darpa.mil/program/systematizing-confidence-in-open-research-and-evidence>
- Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37(1), 60–82.
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science (New York, N.Y.)*, 330(6004), 686–688.
- Wurman, P. R., Wellman, M. P., & Walsh, W. E. (2001). A parametrization of the auction design space. *Games and Economic Behavior*, 35(1), 304–338.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, E1. <https://doi.org/10.1017/S0140525X20001685>
- Yarkoni, T., Eckles, D., Heathers, J., Levenstein, M., Smaldino, P. E., & Lane, J. I. (2019). *Enhancing and accelerating social science via automation: Challenges and opportunities*. <https://doi.org/10.31235/osf.io/vncwv>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12(6), 1100–1122.
- Zelditch, M., Jr. (1969). Can you really study an army in the laboratory. *A Sociological Reader on Complex Organizations*, 528–539.

## Open Peer Commentary

### Integrative experiments require a shared theoretical and methodological basis

Pietro Amerio<sup>a\*</sup> , Nicolas Coucke<sup>a,b</sup>   
and Axel Cleeremans<sup>a</sup> 

<sup>a</sup>Consciousness Cognition and Computation Group, Center for Research in Cognition & Neurosciences, Université Libre de Bruxelles, Brussels, Belgium and <sup>b</sup>IRIDIA, Université Libre de Bruxelles, Brussels, Belgium

[pietro.amerio@ulb.be](mailto:pietro.amerio@ulb.be)  
[nicolas.coucke@ulb.be](mailto:nicolas.coucke@ulb.be)  
[axel.cleeremans@ulb.be](mailto:axel.cleeremans@ulb.be)  
<https://axc.ulb.be/>

\*Corresponding author.

doi:10.1017/S0140525X2300225X, e34

**Abstract**

Creating an integrated design space can be successful only if researchers agree on how to define and measure a certain phenomenon of interest. Adversarial collaborations and mathematical modeling can aid in reaching the necessary level of agreement when researchers depart from different theoretical perspectives.

We agree with Almaatouq et al.'s target article that there is a need for addressing the incommensurability of behavioral experiments and we support the proposed integrative design framework. However, we would like to highlight that the incommensurability of experimental results might stem not only from differences in experimental conditions or populations, but also from disagreements on how to define and measure the phenomenon of interest. While reaching a consensus is not strictly necessary for research to progress, we argue that the success of the integrative approach critically depends on finding agreed-upon theoretical and methodological frameworks.

As an analogy, imagine two researchers, Scarlett and Amber, who study the phenomenon of "pinkness." Scarlett uses a design space that has three dimensions, corresponding to the three base colors of the RGB system (i.e., red, green, and blue). After experimenting with various color combinations, she identifies the area of RGB space in which the color pink is produced. Amber, however, defined her experiments in the YMCK color space. How can Scarlett and Amber's experiments be integrated in a single design space? Two conditions should be met. First, the definition of what "pink" is must be shared between the two scientists. If the range of colors that Amber classifies as "pink" is wider than Scarlett's, then mapping the results of their experiments onto each other is meaningless. Once there is agreement on the definition of pinkness, the second condition is to have a means of translating the ranges under which the phenomenon occurs from RGB space to YMCK space. Such a translation is quite straightforward in our analogy, but it can become a lot more complex when real experimental paradigms are involved.

Our main point here is that the integrative approach can only be successful when researchers agree on the definition of their phenomenon of interest. This is crucial because the definition affects experiment design. As a concrete example, consider perceptual awareness studies, where different researchers have used many different measures of awareness (Timmermans & Cleeremans, 2015). A recent meta-analysis by Yaron, Melloni, Pitts, and Mudrik (2022) found a clear association between the methodological design of an experiment and the theory of consciousness favored by the researchers: An algorithm could even predict which theory an experiment was testing based on its methodology alone! A central issue in this literature concerns whether awareness of a sensory stimulus should be measured subjectively (i.e., via explicit reports from the participants) or objectively (i.e., as performance in a forced-choice task). Crucially, the two methods rest on different definitions of awareness. Objective measures assume that participants can discriminate stimuli correctly only if they are aware of them, while subjective measures rest on the assumption that awareness can diverge from discrimination performance. The difference is not trivial because it forces researchers to adopt substantially different experimental strategies. When testing for the existence of unconscious perception, for example, subjective approaches relate

explicit report to discrimination performance, while objective approaches compare discrimination performance to implicit measures of perceptual processing, such as reaction times (e.g., Dehaene et al., 1998). Integrating these two research lines in a common experimental space would be unsuccessful because, depending on the definition of awareness one adopts, the task design, the collected measures, and the interpretation of results will differ.

One could attempt to circumvent the problem by placing the two approaches into one large design space and connecting them along an additional dimension representing the "measurement method". However, this strategy is only feasible when the measures share the same theoretical basis. As mentioned in the target article, the design space should reveal the conditions under which a phenomenon emerges. To fulfill this function, it is crucial that the phenomenon of interest is precisely defined. Experiments that rest on opposing theoretical views are generally aimed at detecting (slightly) differently defined phenomena. As such, forcing them in a common design space means building a space in which the effect of a particular range of parameters remains ambiguous. In addition, even when the definition of the phenomenon is agreed upon, each experimental paradigm comes with a set of paradigm-specific parameters. Thus, their integration would require a method of mapping different design spaces onto each other, which might not be straightforward. Below, we suggest two potential tools for resolving such conundrums.

The first is adversarial collaborations. These initiatives bring together researchers with contrasting theoretical views and motivate them to design experiments that directly test one theory against another (Cleeremans, 2022; Melloni, Mudrik, Pitts, & Koch, 2021). Such approaches are currently flourishing in consciousness research (e.g., Melloni et al., 2023). As discussed above, the definition of the phenomenon of interest (i.e., its theoretical basis) is crucial for constructing the design space. By testing predictions of different theories against one another, adversarial collaborations can help researchers decide on one definition around which to build (and explore) a full design space. On a parallel line, adversarial collaborations can result in agreed upon methods to map theoretical frameworks onto one another.

The second tool we recommend is mathematical modeling, which can be particularly helpful when results from different experimental paradigms need to be related. The shape of the design space is specific to the paradigm and relating spaces with different shapes is not always possible. Modeling helps in this task by creating a shared analysis space in which results obtained via different measures can be juxtaposed. We draw another example from perceptual awareness research: King and Dehaene (2014) were able to juxtapose results from six major lines of experiments by constructing an overarching mathematical framework, in which results stemming from unconnectable design spaces can be directly compared.

In conclusion, while strongly supporting the integrative experiments approach, our commentary highlights how it might not be possible to reconcile experiments that adopt different theoretical views on the effect of interest. As such, the usefulness of the approach might depend on the researchers' agreement upon adequate measures and theories. When lacking, tools like adversarial collaborations and mathematical models can help constructing a common design space or connecting otherwise isolated spaces.



**Financial support.** P. A. was supported by an F.R.S.-FNRS Research Project T003821F (40003221) to Axel Cleeremans. N. C. was supported by the program of Concerted Research Actions (ARC) of the Université Libre de Bruxelles. A. C. is a research director with the F.R.S.-FNRS (Belgium).

**Competing interest.** None.

## References

- Cleeremans, A. (2022). Theory as adversarial collaboration. *Nature Human Behaviour*, 6(4), 485–486. <https://doi.org/10.1038/s41562-021-01285-4>
- Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., ... Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395(6702), 597–600. <https://doi.org/10.1038/26967>
- King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1641), 20130204. <https://doi.org/10.1098/rstb.2013.0204>
- Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., ... Tononi, G. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLoS ONE*, 18(2), e0268577. <https://doi.org/10.1371/journal.pone.0268577>
- Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science (New York, N.Y.)*, 372(6545), 911–912. <https://doi.org/10.1126/science.abj3259>
- Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In M. Overgaard (Ed.), *Behavioral methods in consciousness research* (pp. 21–46). Oxford University Press. <https://doi.org/10.1093/acprofoso/9780199688890.003.0003>
- Yaron, L., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour*, 6(4), 593–604. <https://doi.org/10.1038/s41562-021-01284-5>

## The elephant's other legs: What some sciences actually do

Jonathan Baron\* 

Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA  
[jonathanbaron7@gmail.com](mailto:jonathanbaron7@gmail.com)  
<https://www.sas.upenn.edu/baron>

\*Corresponding author.

doi:10.1017/S0140525X23002108, e35

### Abstract

Integrative experiments, as described, seem blindly empirical, as if the question of generality of effects could not be understood through controlled one-at-a-time experiments. But current research using such experiments, especially applied research, can resolve issues and make progress through understanding of cause–effect pathways, leaving to engineers the task of translating this understanding into practice.

Almaatouq et al. claim that the sciences of interest are about theory development. I'm not sure what “theory” means to them. Possibly the term refers to something like a prediction that X will affect Y. The term could also refer to a causal explanation, for example, “people tend to favor harms of omission over harms of action because they think of omission as default and use a heuristic of favoring default” or “because they attend more to actions.” The explanations are usually chains of events,

some of which are mental. Questions about causal explanations are often addressed in the one-at-a-time paradigm, with careful controls, and some of these questions are answered (e.g., Baron & Ritov, 1994).

The one-at-a-time paradigm largely concerns the existence of effects, not their generality (Baron, 2010). Often the subject population of interest is “human beings,” and, I hope, most of them are not yet born. In such cases, all samples are convenience samples. We can do small tests of generality of causal effects. Usually they generalize pretty well, in direction if not in magnitude. Once we know that a causal process exists, we can ask further questions about it to understand it better. Anything we learn through these experiments will probably apply to the process when it exists. But existence, of effects and causal chains, is all we can ever learn from experiments, including the integrative experiments proposed.

Some of these integrative experiments discussed pit causal effects against each other. Because the magnitude of effects depends on all sorts of things, it is not clear that we can conclude with much generality which one wins. Indeed, group synergy might work one way in some situations and the opposite in other situations, but it is not clear that the sample space is sufficient to test all possibilities, and, moreover the mere finding that effects are present in some part of the space does not tell us why. The results seem blindly empirical. By contrast, the one-at-a-time approach, when properly applied, can increase our understanding of how things work.

Even when causal effects do not compete, estimation of their relative magnitude will depend on the space of possibilities (as well as who the subjects are, as the article notes). For example, moral judgments about autonomous vehicles may yield quite different results from moral judgments about bioethics.

A different sense of the term “theory” refers to explanations that tie together diverse phenomena that might at first seem to be unrelated. Freud's theory of unconscious motivation (still at work behind the scenes in social psychology, despite its disappearance from most textbooks) was an example. Other theories are more limited in what they explain, such as the idea that many errors result from substitution of judgments of one attribute for judgments about another, which is usually correlated with the first (Kahneman & Frederick, 2002; also Baron, 1973). In some of these theories, the claim is that something happens in many cases, but we do not know which ones. It is something to look for. Similarly for the “germ theory of disease,” which says that, in trying to find out the cause of a disease, it is a good idea to look for very small organisms. In such cases, integrative design competes with the alternative approach of exploring more examples, such as diseases caused by toxicity or genetic abnormalities. Brute-force empiricism would be unlikely to discover or explain such cases.

In applied sciences, such as medicine, broad theory can help, as the “germ theory of disease,” but this sort of theory, for the most part, is neither absolute nor completely general, unlike what physics tries to do. Most medical research is about the etiology and treatment of particular disorders, one-at-a-time, although sometimes a discovery can apply to several similar disorders.

Examples abound in psychology of increased understanding that results from analysis of particular applied problems. A great deal of modern social psychology arose historically out of attempts to understand the rise of fascism. Some of the cognitive psychology of attention and vigilance arose from the study of



radar operators in World War II (Garner, 1972). Recent research on judgment arose out of attempts to measure the nonmarket value of the harm caused by the Exxon-Valdez oil spill (Kahneman, Ritov, Jacowitz, & Grant, 1993). Research on forecasting was spurred by attempts to understand the failures of intelligence agencies (Dhami, Mandel, Mellers, & Tetlock, 2015). Research on risk perception was provoked by perceived over- and under-regulation of risk (Breyer, 1993; Slovic, 1987). In these cases and many others, we have learned a lot. Sometimes institutions have even changed their decision-making procedures in response to what we have learned.

Applied research in medicine and psychology often involves experimental understanding of phenomena such as disorders or biases. Such understanding informs the efforts of engineers (in the broad sense that includes designers of administrative procedures, decision procedures, systems of psychotherapy, and human-machine interfaces). Engineers try to get things to work by a cycle of build-test-build-test and so forth. The practice of decision analysis, for example, has built on laboratory results such as those concerning the difficulty of assigning weights to attributes (von Winterfeldt & Edwards, 1986). Similar relations between basic one-at-a-time research and application are the work on “nudges” (Thaler & Sunstein, 2008), cognitive behavior therapy (Beck, 1979), forecasting (Dhami et al., 2015), and literacy (Treiman, 1992). Often, as in the last two cases, ultimate applications run into political or institutional resistance.


This sort of research is not based on data alone but also on understanding of what kinds of causal links are plausible. Such understanding often comes from background knowledge from a variety of fields, including (for psychology) philosophy, linguistics, computer science, biology, and politics. Understanding of a phenomenon neither comes from blindly empirical research, nor even from careful controlled experiments uninformed by background knowledge.

**Competing interest.** None.

## References

- Baron, J. (1973). Semantic components and conceptual development. *Cognition*, 2, 189–207.
- Baron, J. (2010). Looking at individual subjects in research on judgment and decision making (or anything). *Acta Psychologica Sinica*, 42, 1–11.
- Baron, J., & Ritov, I. (1994). Reference points and omission bias. *Organizational Behavior and Human Decision Processes*, 59, 475–498.
- Beck, A. T. (Ed.). (1979). *Cognitive therapy of depression*. Guilford Press.
- Breyer, S. (1993). *Breaking the vicious circle: Toward effective risk regulation*. Harvard University Press.
- Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science*, 10(6), 753–757.
- Garner, W. R. (1972). The acquisition and application of knowledge: A symbiotic relation. *American Psychologist*, 27(10), 941–946.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.
- Kahneman, D., Ritov, I., Jacowitz, K. E., & Grant, P. (1993). Stated willingness to pay for public goods: A psychological perspective. *Psychological Science*, 4(5), 310–315.
- Slovic, P. (1987). Perception of risk. *Science (New York, N.Y.)*, 236, 280–285.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Treiman, R. (1992). *Beginning to spell: A study of first-grade children*. Oxford University Press.
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press.

## Assume a can opener

Cory J. Clark<sup>a,b\*</sup> , Calvin Isch<sup>c</sup>, Paul Connor<sup>b</sup>  
and Philip E. Tetlock<sup>a,b</sup>

<sup>a</sup>The Wharton School, University of Pennsylvania, Philadelphia, PA, USA;

<sup>b</sup>School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA and <sup>c</sup>Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA

cjclark@sas.upenn.edu; calvin.isch@gmail.com; paulrobertconnor@gmail.com; tetlock@wharton.upenn.edu

<https://www.coryjclark.com>

<https://www.asc.upenn.edu/people/graduate-student/calvin-isch>

<https://www.paulconnorpsych.com>

<https://www.sas.upenn.edu/tetlock/>

\*Corresponding author.

doi:10.1017/S0140525X2300239X, e36

### Abstract

We propose a friendly amendment to integrative experiment design (IED), adversarial-collaboration IED, that incentivizes research teams from competing theoretical perspectives to identify zones of the design space where they possess an explanatory edge. This amendment is especially critical in debates that have high policy stakes and carry a strong normative-political charge that might otherwise prevent free exchange of ideas.

We lift our title from a joke dating back at least 50 years (Boulding, 1970): A physicist, a chemist, and an economist are stranded on a desert island with only a can of food. The physicist and chemist devise ingenious, discipline-grounded solutions for opening the can: Heat, pressure, force. But their high-abstraction colleague sees a better approach, “Let’s think this through systematically – and start by assuming we have a can opener.”

The story comes to mind because we see integrative experiment design (IED) as a high-abstraction idea that is attractive in principle but that will be difficult to put to practical use. We see a tacit assumption underlying IED – namely, behavioral science is a rigorously self-correcting epistemic community regulated by CUDOS norms of science: Communal data-sharing, universalism, disinterestedness, and organized skepticism (Merton, 1942/1973). In this optimistic view, the central obstacle to making behavioral science far more cumulative is essentially organizational. Investigators are too individualistic, insisting on pursuing their own trademark concepts and methods, which entails setting up false binary oppositions (playing 20 questions with nature) from which their side emerges victorious. If only we could subordinate rambunctious scientific egos to the greater epistemic good – integrative experiment design – rapid progress would follow.

We agree with the authors’ criticisms that behavioral science suffers from a validity crisis (in our view, a more devastating problem for behavioral scientists’ collective credibility than the better-known replication crisis). There are countless contradictory claims in the literature and no means of reconciling them because different research teams rely on different methods to study similar phenomena (Clark, Costello, Mitchell, & Tetlock, 2022). Excessive individualism, however, is not the only problem; excessive conformity is as well. Truly thorough “research cartography,” or

mapping out a comprehensive design space for a phenomenon requires investigators to engage with theories that seem to contradict their own previously published work, with variables that fall far outside their area of expertise, and with deeply dissonant possibilities (Tetlock, 1994; Tetlock, Kristel, Elson, Green, & Lerner, 2000). But for a variety of personal, social, theoretical, and ideological reasons, investigators often balk at even considering certain categories of hypotheses (Clark & Winegard, 2020).

These pockets of collective closed-mindedness will bias – sometimes severely bias – the design space. Consider studies of poverty or educational attainment in which many investigators are unwilling to consider behavioral and genetic explanations (Harden, 2021), studies of gender differences in which evolutionary hypotheses are taboo (Buss & von Hippel, 2018), or studies of team dynamics in which investigators are reluctant to report results that cast doubt on the benefits of demographic diversity (Clark et al., 2023; Eagly, 2016). The list of “off limits” – yet perfectly plausible – explanations for many of the most societally important topics in behavioral science is long. And these tend to be the precise topics where scholars are most at loggerheads and thus most in need of progress.

Unlike cartography of the physical world, abstract spaces in social science cannot always be clearly identified and measured, and this ambiguity makes it easy for IED teams to leave out the strongest challenges to their pet theories and ignore socially costly hypotheses. This is especially true if IED teams are relatively homogenous in their theoretical orientations.

This is a challenging problem, and we doubt any big idea will solve it. However, not to let perfection be the enemy of improvement, we propose that IED will be most productive in the context of adversarial collaborations, in which teams of collaborators include scholars who have previously published from multiple competing theoretical perspectives (Clark & Tetlock, 2023; Kahneman, 2003). Traditionally, adversarial collaborations include pairs of disagreeing scholars (e.g., Abele, Ellemers, Fiske, Koch, & Yzerbyt, 2021; Killingsworth, Kahneman, & Mellers, 2023; Mellers, Hertwig, & Kahneman, 2001), but adversarial-collaboration IEDs could include scholars from multiple or even dozens of formerly competing perspectives who study similar phenomena, such as poverty, educational attainment, or violence.

An adversarial approach helps address the problem of motivation. Many scholars are ambitious and *want* their scientific contributions to be distinctive, novel, important, and widely generalizable, and consequently, they lack the motivation to articulate a thorough design space. Indeed, scholars are often aware of alternative but equally relevant independent variables, dependent variables, and contexts from those they routinely test to support their theories, and they choose to avoid or file drawer these alternative approaches. Requiring scholars to work with theoretical adversaries would increase the likelihood that the research design space includes relevant parameters that might be rejected by or simply unknown to a team of theoretically homogeneous scholars. This would also help *narrow* the design space to the most relevant and high-quality parameters by eliminating those that a subset of the research team considers fatally flawed (thus increasing the feasibility of IEDs).

Additionally, our proposed approach could normalize explicit consideration of taboo and other alternative explanations and explicit inclusion of scholars who forward alternative conclusions. Adversarial-collaboration IEDs may be considered incomplete, or lopsided, or biased without relatively exhaustive sampling from

relevant parameters and the scholars with expertise in those parameters. Although current norms of science threaten scholars with ostracism and other social sanctions for considering alternative conclusions and affiliating with the scholars who forward them, adversarial IEDs could require, and thus incentivize this more disinterested and sedulous approach to scholarship (Nemeth, Brown, & Rogers, 2001).


Adversarial IEDs may also motivate the search for genuine metatheories that can explain apparent discrepancies between leading scholars' preferred theories. Rather than pitting seemingly contradictory hypotheses against one another in a “winner takes all” model of science (e.g., are political rightists more cognitively rigid than leftists *or* is cognitive rigidity symmetrical?), adversarial IEDs could lead to the development of metatheories that explain *in which contexts* different claims are true (Bowes et al., 2023). Over time, this could contribute to a more cooperative (and less acrimonious) scientific environment, in which intellectual adversaries are viewed less as enemies to be demolished than as colleagues in pursuit of truth.

**Competing interest.** None.

## References

- Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*, 128(2), 290–314.
- Boulding, K. E. (1970). *Economics as a science*. McGraw Hill.
- Bowes, S., Clark, C. J., Conway, L. G. III, Costello, T. H., Osborne, D., Tetlock, P., & van Prooijen, J. (2023). An adversarial collaboration on the rigidity-of-the-right, rigidity-of-extremes, or symmetry: The answer depends on the question. <https://doi.org/10.31234/osf.io/4wmx2>
- Buss, D. M., & von Hippel, W. (2018). Psychological barriers to evolutionary psychology: Ideological bias and coalitional adaptations. *Archives of Scientific Psychology*, 6(1), 148–158.
- Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022). Keep your enemies close: Adversarial collaborations will improve behavioral science. *Journal of Applied Research in Memory and Cognition*, 11(1), 1–18.
- Clark, C. J., Fjeldmark, M., Lu, L., Baumeister, R. F., Ceci, S., German, K., ... Tetlock, P. E. (2023, February 24). Taboos and self-censorship among psychology professors (Conference presentation). Society for Open Inquiry in Behavioral Science, Behavioral Science Speakeasy, Atlanta, GA, USA.
- Clark, C. J., & Tetlock, P. E. (2023). Adversarial collaboration: The next science reform. In C. L. Frisby, R. E. Redding, W. T. O'Donohue, & S. O. Lilienfeld (Eds.), *Ideological and political bias in psychology: Nature, scope, and solutions*. Springer.
- Clark, C. J., & Winegard, B. M. (2020). Tribalism in war and peace: The nature and evolution of ideological epistemology and its significance for modern social science. *Psychological Inquiry*, 31(1), 1–22.
- Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance?. *Journal of Social Issues*, 72(1), 199–222.
- Harden, K. P. (2021). *The genetic lottery: Why DNA matters for social equality*. Princeton University Press.
- Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, 58(9), 723–730.
- Killingsworth, M. A., Kahneman, D., & Mellers, B. (2023). Income and emotional well-being: A conflict resolved. *Proceedings of the National Academy of Sciences of the United States of America*, 120(10), e2208661120.
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12(4), 269–275.
- Merton, R. K. (1942/1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.
- Nemeth, C., Brown, K., & Rogers, J. (2001). Devil's advocate versus authentic dissent: Stimulating quantity and quality. *European Journal of Social Psychology*, 31(6), 707–720.
- Tetlock, P. E. (1994). Political psychology or politicized psychology: Is the road to scientific hell paved with good moral intentions?. *Political Psychology*, 15, 509–529.
- Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78(5), 853–870.

## Test many theories in many ways

Wilson Cyrus-Lai<sup>a\*</sup> , Warren Tierney<sup>b</sup>  
and Eric Luis Uhlmann<sup>a</sup>

<sup>a</sup>Organisational Behaviour Area, INSEAD, Singapore, Singapore and

<sup>b</sup>Organisational Behaviour Area/Marketing Area, INSEAD, Singapore, Singapore  
wilson-cyrus.lai@insead.edu  
warren.tierney@insead.edu  
eric.luis.uhlmann@gmail.com

\*Corresponding author.

doi:10.1017/S0140525X23002376, e37

### Abstract

Demonstrating the limitations of the one-at-a-time approach, crowd initiatives reveal the surprisingly powerful role of analytic and design choices in shaping scientific results. At the same time, cross-cultural variability in effects is far below the levels initially expected. This highlights the value of “medium” science, leveraging diverse stimulus sets and extensive robustness checks to achieve integrative tests of competing theories.

Almaatouq et al. argue that the “one-at-a-time” approach to scientific research has led to collections of atomized findings of unclear relevance to each other. They advocate for an integrative approach in which stimuli are varied systematically across theoretically important dimensions. This allows for strong inferences (Platt, 1964) regarding which theory holds the most explanatory power across diverse contexts, as well as the identification of meaningful moderators.

Our research group has addressed this challenge by examining the analytic and design choices that naturalistically emerge across independent investigators as well as the implications for the empirical results (Landy et al., 2020; Schweinsberg et al., 2021; Silberzahn et al., 2018). These crowdsourced many analysts and many design initiatives reveal dramatic dispersion in estimates due to researcher choices, empirically demonstrating the limitations of the one-at-a-time approach (see also Baribault et al., 2018; Botvinik-Nezer et al., 2020; Breznau et al., 2022; Menkveld et al., 2023). At the same time, we have sought to further increase the already high-theoretical value of replications by leveraging them for competitive theory testing. Rather than test the original theory against the null hypothesis, we include new conditions and measures allowing us to simultaneously examine the preregistered predictions of different theoretical accounts (Tierney et al., 2020, 2021). In this manner, we can start to prune the dense theoretical landscape (Leavitt, Mitchell, & Peterson, 2010) found in areas of inquiry characterized by many atomized findings and narrow theories.

In contrast, a striking and unexpected *lack* of variability has emerged in the results when many laboratories collect data using the same methods. In such crowd replication initiatives, cross-site heterogeneity in estimates is far below what one would expect based on intuition and theory (Olsson-Collentine, Wicherts, & van Assen, 2020). From a perspectivist (McGuire, 1973) standpoint, psychological phenomena should emerge in some contexts and be nonexistent or even reversed in others (see also Henrich, Heine, & Norenzayan, 2010). And yet, effects

seem to either fail to replicate across all populations sampled or emerge again and again (see also Delios et al., 2022).

Bringing many designs, analyses, theories, and data collection teams together, we recently completed a crowdsourced initiative that qualifies as the type of comprehensive integrative test that Almaatouq et al. envision. Tierney et al. (2023) systematically re-examined the relationships between anger expression, target gender, and status conferral. In the original research, women who displayed anger in professional settings suffered steep drops in the status and respect they were accorded by social perceivers (Brescoll & Uhlmann, 2008). In the original investigations, only a single set of videos featuring one female and one male target were employed as stimuli, and all participants were from Connecticut. In contrast, the crowdsourced replication project featured 27 experimental designs, a multiverse capturing many defensible analytic approaches, and 68 data collection sites in 23 countries. We further tested the original prescriptive stereotype account against competing theories predicting that anger signals status similarly for women and men, that anger has vastly different status implications in Eastern and Western cultures, and that feminist messaging has successfully reduced or even reversed gender biases. As Almaatouq et al. recommend, we probed the dose-response relationship between anger and status conferral by both experimentally manipulating and measuring the extremity of emotion expressions across different designs.

The crowd initiative finds that anger increases status by signaling dominance and assertiveness, while also diminishing it by projecting incompetence and unlikability, aggregating across a wide range of research approaches and populations. Critically, this same pattern emerged for both female and male targets, social perceivers of different genders, and in both Eastern and harmony-oriented cultures and Western and more conflict-oriented ones. Highlighting the value of deploying diverse research approaches, six of the 27 designs found favoritism toward men in status conferral, but one design pointed to the opposite conclusion. Similarly, in a multiverse with 32 branches, there existed just two specifications that supported the original gender-and-anger backlash effect. Had we employed a one-at-a-time approach, we could have accidentally hit upon or strategically chosen narrow methods yielding nonrepresentative conclusions (e.g., of profemale status bias or gender backlash). Overall, the intellectual returns on including many designs, many analyses, and many theories were high. In contrast, and consistent with past crowd initiatives, collecting data across many places revealed minimal cross-site heterogeneity and no interesting cultural differences.

Thus, we envision a diverse scientific ecology consisting of many “small” and “medium” projects and just a few huge international efforts. The one-at-a-time approach is an efficient means to introduce initial evidence for promising new hypotheses. However, as a theoretical space becomes increasingly cluttered, intellectual returns are maximized by sampling stimuli widely and employing many analyses to provide severe tests of competing theories (Mayo, 2018). Although this could involve a crowd of laboratories, a single team could carry out a multiverse (Steege, Tuerlinckx, Gelman, & Vanpaemel, 2016) and operationalize key variables in a variety of ways. A small team might sample just one or two participant populations that are easily accessible to them. Finally, a subset of findings of particularly high-theoretical and practical importance should be selected for crowdsourced data collections across many nations as a systematic test of cross-cultural generalizability. When numerous sites are not available, the researchers might carry out the first



generalizability test in the most culturally distant population available (Muthukrishna et al., 2020). If the effect is still observed, this represents initial evidence of universality (Norenzayan & Heine, 2005).

In sum, an ironic legacy of the movement to crowdsource behavioral research may be showing that scaling science to such a massive level might be neither efficient nor strictly necessary for most research findings. The sorts of integrative tests Almaatouq et al. envision can also be accomplished by a small team that actively ensures a diversity of analyses and stimuli, and yet collects data locally or across a few carefully selected cultures rather than globally. In the future, our greatest intellectual returns on investment may come from “medium” science that prioritizes testing many theories in many ways.

**Financial support.** This research was supported by an R&D grant from INSEAD to Eric Luis Uhlmann.

**Competing interest.** None.

## References

- Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2607–2612.
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582, 84–88.
- Brescoll, V. L., & Uhlmann, E. L. (2008). Can an angry woman get ahead? Status conferral, gender, and expression of emotion in the workplace. *Psychological Science*, 19, 268–275.
- Brezna, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, 119(44), e2203150119.
- Delios, A., Clemente, E., Wu, T., Tan, H., Wang, Y., Gordon, M., ... Uhlmann, E. L. (2022). Examining the context sensitivity of research findings from archival data. *Proceedings of the National Academy of Sciences of the United States of America*, 119(30), e2120377119.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83.
- Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5), 451–479.
- Leavitt, K., Mitchell, T., & Peterson, J. (2010). Theory pruning: Strategies for reducing our dense theoretical landscape. *Organizational Research Methods*, 13, 644–667.
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.
- McGuire, W. J. (1973). The yin and yang of progress in social psychology: Seven koan. *Journal of Personality and Social Psychology*, 26(3), 446–456.
- Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., ... Wu, Z.-X. (2023). Non-standard errors. *The Journal of Finance*. <http://dx.doi.org/10.2139/ssrn.3961574>
- Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, 31(6), 678–701.
- Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin*, 135, 763–784.
- Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, 146(10), 922–940.
- Platt, J. R. (1964). Strong inference. *Science (New York, N.Y.)*, 146, 347–353.
- Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O., van Aert, R., van Assen, M., ... Uhlmann, E. (2021). Radical dispersion of effect size estimates when independent scientists operationalize and test the same hypothesis with the same data. *Organizational Behavior and Human Decision Processes*, 165, 228–249.
- Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtry, E., ... Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, 1, 337–356.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702–712.
- Tierney, W., Cyrus-Lai, W., ... (2023). Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral. Unpublished manuscript.
- Tierney, W., Hardy, J. H. III., Ebersole, C., Leavitt, K., Viganola, D., Clemente, E., ... Uhlmann, E. (2020). Creative destruction in science. *Organizational Behavior and Human Decision Processes*, 161, 291–309.
- Tierney, W., Hardy, J. H. III., Ebersole, C. R., Viganola, D., Clemente, E. G., Gordon, M., ... Uhlmann, E. L. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology*, 93, 104060.

## There are no shortcuts to theory

Berna Devezer\* 

College of Business and Economics, University of Idaho, Moscow, ID, USA

[bdevezer@uidaho.edu](mailto:bdevezer@uidaho.edu)

<https://webpages.uidaho.edu/bernadevezer/>

\*Corresponding author.

doi:10.1017/S0140525X23002169, e38

### Abstract

Almaatouq et al. claim that the integrative experiment design can help “develop a reliable, cohesive, and cumulative theoretical understanding.” I will contest this claim by challenging three underlying assumptions about the nature of scientific theories. I propose that the integrative experiment design should be viewed as an exploratory framework rather than a means to build or evaluate theories.

I contest Almaatouq et al.’s claim that the integrative experiment design can help “develop a reliable, cohesive, and cumulative theoretical understanding” (target article, sect. 3.3, para. 1). This claim relies on three assumptions which I will challenge.

*Assumption 1:* Experiments in social and behavioral sciences test theories.

*Challenge:* Theory tests are, in fact, rare.

The authors assume that statistical null hypothesis tests in social and behavioral sciences involve a theoretical prediction, but this assumption has been widely challenged (e.g., Meehl, 1967, 1978, 1990). Recent scholarship in psychological science suggests that there is a *theory crisis* (e.g., Eronen & Bringmann, 2021; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Oude Maatman, 2021; Proulx & Morey, 2021; Robinaugh, Haslbeck, Ryan, Fried, & Waldorp, 2021; van Rooij, 2019), rather than the “increasing theoretical maturity” (target article, sect. X, para. X) that the target article claims. Perhaps *theoretical amnesia* (Borsboom, 2013) can explain this discord, where researchers can no longer tell what a theory is, and statistical models occupy the vacuum created by the absence of theories. Navarro (2021) and Gelman (2022) warn us against mistaking statistical models and inferences for scientific theories, just as Gigerenzer (1998) did a few decades earlier, pointing out that most statistical hypotheses being tested correspond to misleading surrogates for theory rather than genuine theoretical predictions. Almaatouq et al. provide no evidence to convince us otherwise, and their approach is more likely to integrate/reconcile such surrogates than to make real theoretical progress.

**Assumption 2:** Theories can be meaningfully reduced to a set of conditions represented by a set of experimental parameters.  
**Challenge:** Theories are more than effects and their boundary conditions.

In rare cases where a well-specified scientific theory exists, we need to be explicit about what it entails and how it relates to the experiment. Almaatouq et al. assume that theories can be meaningfully captured by a set of (boundary) conditions that define experimental parameters in the design space. This reductionist view of theory overlooks the role of mechanisms, explanations, and understanding in scientific theory, and overemphasizes the mapping between experimental parameters and theory.

There is no universal consensus on what a theory entails (Winther, 2021) but a rudimentary framework à la Suppes (1967) will serve here. A scientific theory comprises two distinct components: An abstract logical calculus using symbolic representations of a set of propositions, and a set of rules that give the logical calculus empirical content. The formal part of theory is used to *represent*, *explain*, and/or to *predict* an empirical phenomenon (Guest & Martin, 2021). However, a formalism capturing aspects of a phenomenon without offering any mechanistic or causal scientific explanation does not automatically amount to theory (McMullin, 2008; van Rooij & Baggio, 2021). Theories gain explanatory power by isolating the causes and uncovering the mechanism generating empirical regularities (Craver, 2006; Rohrer, 2018). Yet experiments often aim to discover or confirm “effects” signifying empirical facts without providing an explanation. Per Cummins (2010), McGurk effect captures a regularity regarding how speech sounds are perceived across senses, however, it cannot elucidate why the observed regularity occurs. As Poincaré (1905) observes: “Science is built up of facts, as a house is built of stones; but an accumulation of facts is no more a science than a heap of stones is a house.” Scientific theories should go beyond accumulating empirical effects and their boundary conditions to inform us about the structure of systems they purport to explain (Van Rooij & Baggio, 2021).

Using experiments to isolate effects and boundary conditions as a means to test theories would still face the issue of underdetermination of scientific theory by evidence, even if we only needed theories for description and prediction, not explanation (Stanford, 2021). We depend on auxiliary assumptions to derive empirical consequences from a theory, and typically these assumptions neither uniquely identify distinct theories nor remain fixed over time. The integrative design purports to sample the set of experimental parameters comprising such auxiliary assumptions regarding experimental paradigm, context, population of interest, measurements, and so on. In fields where theories are rarely precise enough to specify these assumptions, any experimental design would be conceptually removed from the theory it is meant to test. When theory–experiment mapping is weak, we can define multiple empirically equivalent theories that can reasonably account for the observed data. Alternatively, a given dataset can be used to refute or confirm a given theory simply by altering how auxiliary assumptions are related to the theory.

We cannot simply assume that the design space for an integrative experiment effectively and exclusively captures key features of a theory; extensive theoretical development needs to precede explicit mapping of theory and experiment.

**Assumption 3:** Reconciling inconsistent experimental results may help reconcile incommensurable theories.

**Challenge:** Incommensurability of scientific theories is not an empirical problem.

The target article uses the term “(in)commensurability” (target article, sect. X, para. X) to characterize apparent inconsistencies or incomparabilities in experimental designs and results, suggesting that integrative design can effectively reconcile or compare incommensurable theories. In the philosophy of science, *semantic* incommensurability means there is no common measure between theories, and their fundamental concepts cannot be meaningfully translated or logically related to one another (Oberheim & Hoyningen-Huene, 2018) while *methodological* incommensurability involves unattainability of shared, external, neutral methodological standards to perform a comparative evaluation theories (Chang, 2012). Both kinds of theoretical incommensurability lead to underdetermination of theory choice. Empirical inconsistency described by the authors does not invoke theoretical incommensurability; rather it points to a lack of properly specified models or uncertainty associated with statistical inferences. True theoretical incommensurability cannot be reconciled with an integrative (or any other) experimental design, by definition. It has even been argued that forced reconciliation among incommensurable theories is not desirable and an independent pluralistic existence is necessary for theoretical progress (Chang, 2012).

Integrative experimental design may serve a crucial *exploratory* role in the scientific landscape, by methodically narrowing down experimental conditions that are necessary to observe a phenomenon. Indeed the notion of systematic exploratory experimentation is not new (Burian, 1997; Steinle, 1997) while largely underappreciated. The targeted theoretical aims, however, seem unfeasible if not impossible. There are no empirical shortcuts to theoretical progress.

**Acknowledgments.** I would like to thank Esther Mondragón and Erkan O. Buzbas for their insightful feedback.

**Financial support.** This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (Award No. P20GM104420).

**Competing interest.** None.

## References

- Borsboom, D. (2013, November 20). Theoretical amnesia [Blog]. Open Science Collaboration Blog. <http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/>
- Burian, R. M. (1997). Exploratory experimentation and the role of histochemical techniques in the work of Jean Brachet, 1938–1952. *History and Philosophy of the Life Sciences*, 19(1), 27–45. <https://www.jstor.org/stable/23332033>
- Chang, H. (2012). Incommensurability: Revisiting the chemical revolution. In V. Kindi & T. Arabatzis (Eds.), *Kuhn's the structure of scientific revolutions revisited* (p. 153). Routledge.
- Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3), 355–376. <https://doi.org/10.1007/s11229-006-9097-x>
- Cummins, R. (2010). “How does it work?” vs. “what are the laws?": Two conceptions of psychological explanation. In *The world in the head* (pp. 282–310). Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199548033.003.0016>
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4), 779–788. <https://doi.org/10.1177/1745691620970586>
- Gelman, A. (2022). Mismatch between scientific theories and statistical models. *Behavioral and Brain Sciences*, 45, e15. <https://doi.org/10.1017/S0140525X21000091>
- Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, 8(2), 195–204. <https://doi.org/10.1177/0959354398082006>
- Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16(4), 789–802. <https://doi.org/10.1177/1745691620970585>

- McMullin, E. (2008). The virtues of a good theory. In S. Psillos & M. Curd (Eds.), *The Routledge companion to philosophy of science* (pp. 498–508). Routledge.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115. <https://doi.org/10.1086/288135>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195–244. <https://doi.org/10.2466/pr0.1990.66.1.195>
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3), 221–229. <https://doi.org/10.1038/s41562-018-0522-1>
- Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, 16(4), 707–716. <https://doi.org/10.1177/1745691620974769>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596–1618. <https://doi.org/10.3758/s13423-019-01645-2>
- Oberheim, E., & Hoynigen-Huene, P. (2018). The incommensurability of scientific theories. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2018/entries/incommensurability/>
- Oude Maatman, F. (2021). Psychology's theory crisis, and why formal modelling cannot solve it [Preprint]. *PsyArXiv*. <https://doi.org/10.31234/osf.io/puqvs>
- Poincaré, H. (1905). *Science and hypothesis* (W. J. Greenstreet, Trans.). Walter Scott.
- Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, 16(4), 671–681. <https://doi.org/10.1177/17456916211017098>
- Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, 16(4), 725–743. <https://doi.org/10.1177/1745691620974697>
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Stanford, K. (2021). Underdetermination of scientific theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2021/entries/scientific-underdetermination/>
- Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, 64(S4), S65–S74. <https://doi.org/10.1086/392587>
- Suppes, P. (1967). What is a scientific theory? In S. Morgenbesser (Ed.), *Philosophy of science today* (pp. 55–67). Basic Books.
- van Rooij, I. (2019, January 18). Psychological science needs theory development before preregistration. Psychonomic Society Featured Content. <https://featuredcontent.psychonomic.org/psychological-science-needs-theory-development-before-preregistration/>
- Van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 16(4), 682–697. <https://doi.org/10.1177/1745691620970604>
- Winther, R. G. (2021). The structure of scientific theories. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2021/entries/structure-scientific-theories/>

## Integrative design for thought-experiments

Daniel Dohrn<sup>a\*</sup> and Angelica Mezzadri<sup>b</sup>

<sup>a</sup>Dipartimento di Filosofia “Piero Martinetti,” Università degli Studi di Milano, Milano, Italy and <sup>b</sup>Dipartimento di Filosofia e Scienze dell’Educazione, Università degli Studi di Torino, Torino, Italy  
[daniel.dohrn@unimi.it](mailto:daniel.dohrn@unimi.it)  
[angelica.mezzadri@unito.it](mailto:angelica.mezzadri@unito.it)

<https://www.unimi.it/ugov/person/daniel-dohrn>  
<https://www.finophd.eu/phd-students/161>

\*Corresponding author.

doi:10.1017/S0140525X23002340, e39

### Abstract

Integrative experiment design should be extended to thought-experiments. Thought-experiments are closely connected to “real” experiments. They are involved in devising the design space of theories and possible experiments. The latter may be partitioned into experiments to be really performed and mere thought-experiments. The proposed extension of integrative experiment design lends guidance to a more methodical performance of thought-experiments.

There are four reasons why integrative experiment design as developed by Almaatouq et al. should be extended to thought-experiments.

First, while thought-experiments differ from “real” experiments in being performed only mentally (Kuhn, 1977), there is a close connection: “thought experiment is experiment (albeit a limiting case of it)” (Sorensen, 1992, p. 3). One tempting idea is that a thought-experimenter performs an experiment on herself: She runs her own cognitive processes off-line with a certain hypothetical input and then reports the outcome. For instance, in the moral machine experiment (Awad et al., 2018), participants are faced with a certain decision situation. They take the hypothetical situation as an input for running their decision-making process off-line (without actually taking a decision) and report the outcome.

Second, the practice of thought-experimenting is widespread in science, including the social and behavioural sciences. Here, thought-experiments play a role in producing, justifying, and refuting scientific theories. Typical examples from the social sciences are Keynes’s beauty contest scenario (Kornberger & Mantere, 2020), DuBois’s colour line scenario, or Addams’s scenario where only women are allowed to vote (Hill, 2005). Moral machine experiments started from ethical thought-experiments and extended them. Participants were invited to perform a thought-experiment, and their responses were treated as data in a real social experiment. The widespread use of thought-experiments calls for methodological reflection.

Third, the practice of thought-experimenting aggravates the incommensurability problems with real experiments. The limited description of a thought-experiment leaves many details unspecified. Thought-experiments are performed only occasionally, one after the other, without coordinating them with each other and with theorizing. At most they follow the scheme “question → theory → hypothesis → experiment → analysis → revision to theory → repeat,” as for instance Gettier experiments in epistemology (Gettier, 1963; Praëm & Steglich-Petersen, 2015). They are not systematically controlled for variables like implicit biases of the experimenter. There is no systematic variation in their explicit input. Completeness with respect to the target theory is not prioritized. The trolley case and its numerous variations are a clear example of how these problems affect thought-experiments (Dewitt, Fischhoff, & Salin, 2019; Foot, 1967).

These issues are not sufficiently discussed in the literature. The latter addresses what a thought-experiment is (Sorensen, 1992), how thought-experiments are processed (Nersessian, 2007), whether and how they can provide evidence without empirical observation (Brown, 1991), what their function and scope is (Praëm & Steglich-Petersen, 2015), how they relate to arguments (Norton, 2004), and how to formalize them (Dohrn, 2018; Williamson, 2007). If incommensurability problems are discussed, then mostly with a sceptical twist (Machery, 2017).



What is lacking is a more *constructive methodology for when and how to perform thought-experiments* in a coordinated manner that contributes to building and testing theories.

Fourth, thought-experimenting can be a useful or even indispensable device for integrative experiment design. It arguably plays a role in coming up with such a design, and the perspective on thought-experiments naturally supplements the design space.

Having outlined four reasons for giving thought-experiments a role in integrative experiment design, we distinguish two directions of relevance. Thought-experiments are relevant to integrative experiment design and vice versa.

Integrative experiment design proceeds from a research question to a design space of theories and possible experiments with regards to relevant dimensions of independent variables. Thought-experiments are relevant to this procedure.

First, thought-experiments are heuristically useful in ruling out certain theories and focusing on others. For instance, Galilei's thought-experiments ruled out the Aristotelian theory of motion and oriented scientists towards Newtonian mechanics. Moreover, Almaatouq et al. suggest that *candidate dimensions of relevant variables* are taken "either from the literature or from experience" (target article, sect. 3.1, para. 6). Yet before the advent of integrative design, neither literature nor experiments were guaranteed to systematically survey relevant dimensions. Thought-experimenting provides an efficient heuristic for identifying these dimensions. Take the moral machine experiment: One has to check both variables explicitly fixed (rich vs. poor, old vs. young, etc.) and implicit variables that may have an influence on the decision (nationality, education, religion, etc.). One heuristic step in identifying such variables is a mental simulation of their impact on one's own decision making.

Second, integrative design proceeds by *identifying a universe of possible experiments in a domain of inquiry*. Since not all of these experiments can be really performed, one has to define an order of priority and a coordinating design plan for the experiments to be really performed. Beyond the plan of real experiments, there are those possible experiments in the universe which are dismissed as irrelevant, but there are also those which are relevant but not really performed for various motives: Too complicated, too costly, unethical, or simply too numerous. Sometimes one can reliably anticipate the result of an experiment without having to really perform it. Therefore, among the relevant experiments which are not part of the planned real experiments, one may select those to be performed as thought-experiments. These thought-experiments also should be ordered according to their priority and coordinated with regards to the explicit and implicit parameters to be controlled for.

The last paragraph already shows how integrative experiment design is relevant for the practice of thought-experimenting. Until now, thought-experiments have been largely performed in an anarchic manner. The proposed extension of integrative experiment design lends guidance to *performing them methodically*. It surveys potential thought-experiments, subjects them to an order of priority, and fixes parameters for explicit variation and parameters to be controlled for. It renders experimental settings comparable and reduces problems with incommensurability.

We note in closing that the anarchic mode of performing thought-experiments may sometimes be useful in playing a disruptive role (Stuart, 2020). Such a role may not be captured by integrative design, but integrative design does not exclude it either.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Brown, J. (1991). Thought experiments: A platonic account. In *Thought experiments in science and philosophy* (pp. 119–128). Rowman & Littlefield.
- Dewitt, B., Fischhoff, B., & Salin, N. E. (2019). "Moral machine" experiment is no basis for policymaking. *Nature*, 567(7746), 31. <https://doi.org/10.1038/d41586-019-00766-x>
- Dohrn, D. (2018). Thought experiments without possible worlds. *Philosophical Studies* 175, 363–384. <https://doi.org/10.1007/s11098-017-0871-z>
- Foot, P. (1967). The problem of abortion and the doctrine of double effect. In *Virtues and vices* (pp. 5–15). Oxford University Press.
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123.
- Hill, M. R. (2005). Sociological thought experiments: Five examples from the history of sociology. *Sociological Origins*, 3(2), 3–19.
- Kornberger, M., & Mantere, S. (2020). Thought experiments and philosophy in organizational research. *Organization Theory*, 1(3), 1–19. <https://doi.org/10.1177/2631787720942524>
- Kuhn, T. (1977). *The essential tension*. University of Chicago.
- Machery, E. (2017). *Philosophy within its proper bounds*. Oxford University Press.
- Nersessian, N. J. (2007). Thought experimenting as mental modeling: Empiricism without logic. *Croatian Journal of Philosophy*, 7(20), 125–161.
- Norton, J. D. (2004). On thought experiments: Is there more to the argument? *Philosophy of Science*, 71(5), 1139–1151. <https://doi.org/10.1086/425238>
- Praëm, S. K., & Steglich-Petersen, A. (2015). Philosophical thought experiments as heuristics for theory discovery. *Synthese*, 192(9), 2827–2842. <https://doi.org/10.1007/s11229-015-0684-6>
- Sorensen, R. A. (1992). *Thought experiments*. Oxford University Press.
- Stuart, M. T. (2020). The productive anarchy of scientific imagination. *Philosophy of Science*, 87(5): 968–978. <https://doi.org/10.1086/710629>
- Williamson, T. (2007). *The philosophy of philosophy*. Blackwell.

## Explore your experimental designs and theories before you exploit them!

Marina Dubova<sup>a\*</sup>, Sabina J. Sloman<sup>b</sup>, Ben Andrew<sup>c,d</sup>,  
Matthew R. Nassar<sup>c,d</sup> and Sebastian Musslick<sup>c,d,e</sup>

<sup>a</sup>Cognitive Science Program, Indiana University Bloomington, IN, USA;

<sup>b</sup>Department of Computer Science, University of Manchester, Manchester, UK;

<sup>c</sup>Cognitive, Linguistic, and Psychological Sciences, Brown University,

Providence, RI, USA; <sup>d</sup>Carney Institute for Brain Science, Brown University,

Providence, RI, USA and <sup>e</sup>Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany

[mdubova@iu.edu](mailto:mdubova@iu.edu)

[sabina.sloman@manchester.ac.uk](mailto:sabina.sloman@manchester.ac.uk)

[benjamin\\_andrew@brown.edu](mailto:benjamin_andrew@brown.edu)

[matthew\\_nassar@brown.edu](mailto:matthew_nassar@brown.edu)

[musslick@brown.edu](mailto:musslick@brown.edu)

<https://www.mdubova.com/>

[www.smusslick.com](http://www.smusslick.com)

\*Corresponding author.

doi:10.1017/S0140525X23002303, e40

### Abstract

In many areas of the social and behavioral sciences, the nature of the experiments and theories that best capture the underlying constructs are themselves areas of active inquiry. Integrative experiment design risks being prematurely exploitative, hindering exploration of experimental paradigms and of diverse theoretical accounts for target phenomena.

Almaatouq et al. argue that one-at-a-time experiments hamper efficient exploration of target phenomena and theoretical integration. To address this, they suggest integrative experimentation: Data collection in a large, predetermined, experimental design space. Although integrative experimentation addresses many limitations of current experimental practices in the social and behavioral sciences, we argue that integrative experimentation risks being prematurely exploitative by (a) committing to existing experimental paradigms and dimensions of the corresponding design space, and (b) imposing constraints on theory-building. One-at-a-time experimentation serves a critical role in exploring useful experimental and theoretical paradigms that can then be effectively exploited by integrative experimentation.

### *Integrative experimentation exploits existing experimental paradigms and dimensions of the corresponding design spaces*

Although integrative experimentation facilitates exploration within the prespecified design space, it exploits the information – or lack thereof – that informs the characterization of this space. To perform integrative experiments, scientists must identify a priori a small set of experimental tasks to invest in. Almaatouq et al. present several illustrative examples: Peterson, Bourgin, Agrawal, Reichman, and Griffiths (2021) invested enormous resources to collect human decisions on ~10,000 bandit gambles, Baribault et al. (2018) focused on a specific subliminal priming task, and Awad et al. (2018, 2020) extensively sampled a space of trolley problems. In fields where the nature of the experiments that best measure the underlying constructs are themselves areas of active inquiry, experiments are run under imperfect knowledge about the paradigm that will best capture a target phenomenon. One-at-a-time experimentation enables open-ended, cheap, and sequentially adaptive exploration of experimental paradigms and assumptions about the design spaces corresponding to these paradigms – including exploration along previously unexplored dimensions of a theoretically infinite design space.

Most areas of social and behavioral sciences use experimental manipulations and outcomes to measure unobservable constructs. Social and behavioral scientists in most domains are still engaged in iterative refinement of the experimental paradigms and dimensions of the design space that will best measure these constructs (Dubova & Goldstone, 2023). For instance, while a plethora of paradigms – including the multisource interference task, the task switching paradigm, and the N-back task – are utilized for the study of mental effort, there is little agreement about which experimental manipulations evoke mentally effortful processes, let alone how these manipulations would be combined into an integrative experiment (Bustamante et al., 2022; Koch, Poljac, Müller, & Kiesel, 2018; Kool, McGuire, Rosen, & Botvinick, 2010; Shenhav et al., 2017; Westbrook & Braver, 2015). Here, running integrative experiments can hinder solving the main problem of the field – identifying a set of experimental manipulations relevant to the construct of mental effort.

Almaatouq et al. give examples of areas in the social and behavioral sciences that are dominated by a small set of “standard” experimental paradigms, such as bandit gambles for risky decision making. In these cases, integrative experimentation can facilitate efficient exploration of behavior across the space defined by these paradigms. In other cases, however, integrative experimentation may actually hinder exploration of the target phenomena. For instance, early vision science operated in design spaces involving artificial visual stimuli. While integrative experimentation would have yielded theoretical commensurability in this

space, one-at-a-time experiments (i.e., the use of stimuli that differed from the common design space) enabled a quick expansion of the space to natural stimuli that in turn led to rapid revisions of dominant theories of vision (Olshausen & Field, 2005; Zhaoping, 2014). Thus, scientific inquiry may often not justify a large investment of resources and interinstitutional coordination at the expense of expanding the design space or developing a number of completely new tasks.

### *Integrative experimentation exploits existing theoretical paradigms*

Almaatouq et al. advocate for using integrative experiments to enforce commensurability of theoretical accounts for the data. However, this approach may prematurely prioritize some theoretical frameworks over others. For example, the BrainScore benchmark integrates neuroimaging studies on visual object recognition to standardize the comparison of formal theories of neural visual processing (Schrimpf et al., 2020). Although aiming for inclusivity, BrainScore’s design required certain commitments, such as the set of target phenomena and measurements to be accounted for (i.e., neural recordings in object recognition experiments) and the form that the theories can take (i.e., neural networks mimicking the ventral stream, taking pixels as inputs, and predicting behavioral responses). Equally justified alternative benchmarks could have led to different theories of visual processing being prioritized: For instance, the dataset could have emphasized temporal aspects of vision, or clumped together object recognition with visual search tasks when identifying the domain space for theories to capture. Similarly, standardizing theoretical accounts by the constraints imposed by integrative experiments, which often focus on a single experimental paradigm, may hinder exploration of theoretical frameworks that target different aspects of the phenomena.

Many, if not most, areas of social and behavioral sciences would benefit from facilitating investigation of a larger class of theoretical paradigms, rather than constraining theory-building. For example, cognitive science consists of incommensurable theoretical paradigms, such as rational analysis and dynamical systems, which make predictions about different, often nonoverlapping, aspects of cognitive phenomena. For instance, dynamical systems modeling seeks to capture the temporal aspects of a cognitive process, whereas rational analysis focuses on the outcomes of cognition. A diversity of theoretical frameworks informs the design of new experimental paradigms, broadens the collective conceptualization of the relevant design spaces (Chang, 2012; Massimi, 2022), and contributes to more comprehensive insights on cognition (Krakauer, Ghazanfar, Gomez-Marín, MacIver, & Poeppel, 2017; Marr, 1982; Medin & Bang, 2014). Constraining theory-building risks reinforcing biases that favor existing experimental paradigms, further inhibiting exploration of novel experimental and theoretical frameworks (Dubova, Moskvichev, & Zollman, 2022; Sloman, Oppenheimer, Broomell, & Shalizi, 2022).

### *Integrative and one-at-a-time experimentation benefit fields with different goals at different stages of their development*

Viewed from a resource allocation perspective, scientific endeavors face an explore–exploit dilemma. Integrative experimentation facilitates broad characterization of behavior within a specific paradigm and its corresponding design space. One-at-a-time experimentation encourages iterative refinement of experimental paradigms and the development of new theoretical frameworks. We believe a combination of

integrative and one-at-a-time experimentation is needed to effectively address the explore–exploit problem in sciences.

**Financial support.** S. M. is supported by Schmidt Science Fellows, in partnership with the Rhodes Trust. M. R. N. is supported by NIH RO1 MH126971.

**Competing interest.** None.

## References

- Awad, E., Dsouza, S., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, 63(3), 48–55.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64.
- Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2607–2612.
- Bustamante, L. A., Oshinowo, T., Lee, J. R., Tong, E., Burton, A. R., Shenhav, A. S., ... Daw, N. D. (2022). Effort foraging task reveals positive correlation between individual differences in the cost of cognitive and physical effort in humans and relationship to self-reported motivation and affect. *bioRxiv*, 2022-11.
- Chang, H. (2012). *Is water H<sub>2</sub>O?: Evidence, realism and pluralism* (Vol. 293). Springer Science & Business Media.
- Dubova, M., & Goldstone, R. L. (2023). Carving joints into nature: reengineering scientific concepts in light of concept-laden evidence. *Trends in Cognitive Sciences*, 27(7), 656–670.
- Dubova, M., Moskvichev, A., & Zollman, K. (2022). Against theory-motivated experimentation in science. *MetaArXiv*, June, 24.
- Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking – An integrative review of dual-task and task-switching research. *Psychological Bulletin*, 144(6), 557.
- Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, 139(4), 665.
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marín, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3), 480–490.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.
- Massimi, M. (2022). *Perspectival realism*. Oxford University Press.
- Medin, D. L., & Bang, M. (2014). *Who's asking?: Native science, western science, and science education*. MIT Press.
- Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1?. *Neural Computation*, 17(8), 1665–1699.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, 372(6547), 1209–1214.
- Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3), 413–423.
- Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 40, 99–124.
- Slooman, S. J., Oppenheimer, D. M., Broomell, S. B., & Shalizi, C. R. (2022). Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias. *arXiv preprint arXiv:2205.13698*.
- Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, 15, 395–415.
- Zhaoqing, L. (2014). *Understanding vision: Theory, models, and data*. Oxford University Press.

## Confidence in research findings depends on theory

David Gal<sup>a\*</sup>, Brian Sterntal<sup>b</sup> and Bobby J. Calder<sup>b</sup>

<sup>a</sup>Department of Marketing, University of Illinois Chicago, Chicago, IL, USA and

<sup>b</sup>Department of Marketing, Northwestern University, Evanston, IL, USA  
[davidgal@uic.edu](mailto:davidgal@uic.edu); [bst047@kellogg.northwestern.edu](mailto:bst047@kellogg.northwestern.edu); [calder@kellogg.northwestern.edu](mailto:calder@kellogg.northwestern.edu)

\*Corresponding author.

doi:10.1017/S0140525X23002261, e41

## Abstract

Almaatouq et al. view the purpose of research is to map variable-to-variable relationships (e.g., the effect of X on Y). They also view theory as this mapping of variable-to-variable relationships rather than an explanation of why the relationships occur. However, it is theory as explanation that allows us to reconcile disparate findings and that should guide application.

We agree with Almaatouq et al. that the integration of disparate research findings is often inefficient or fails to occur entirely. However, their proposed solution is based on an inadequate but widely shared conception of the nature of theory and its importance for application.

Almaatouq et al. *assume* the aim of research is to map variable-to-variable relationships (e.g., the effect of X on Y), and that current research has failed to do so adequately due to experiments that use incommensurate variables. Their approach is to use the literature and experience including previous experiments to identify a large number of variables that form a “design space” of experimental outcomes. By means such as sampling and predicting outcomes, boundaries in this space can be established that specify disparate sets of outcomes. This process results in a “theory” of how variables work together in complex ways. Almaatouq et al. remark that their integrative approach may strike many as atheoretical. In fact, it is atheoretical in that the focus is entirely on variables as opposed to explanatory theoretical constructs.

For Almaatouq et al. theory is a mapping of variable-to-variable relationships. Their approach entirely ignores the need for an explanation of *why* the relationships occur. Progress in research requires both the observation of variable relationships and their explanation. Observation informs theory and theory informs observation. Neither is sufficient, no matter what the scale of observation.

To illustrate their approach, Almaatouq et al. discuss the phenomenon of group “synergy.” They write that research on the topic often reaches conflicting conclusions, with some studies finding that groups outperform individuals and other studies finding that individuals outperform groups. They lament that, “researchers in this space have no way to articulate how similar or different their experiment is from anyone else’s. As a result, it is impossible to determine... how all of the potentially relevant factors jointly determine group synergy...” (target article, sect. 1, para. 4).

However, the idea that research can determine how “all of the potentially relevant factors” influence the effect of one variable on another is ill-founded. Effects are always contingent, and the moderating variables that influence them are unbounded and evolving. Attempting to circumscribe an effect is a vain pursuit. Only theoretical explanation can resolve this problem.

Consider Almaatouq et al.’s example of the sort of systematic examination involving “commensurate” studies they favor, an investigation of moral dilemmas, “inspired by the trolley problem,” by Awad et al. (2018). They write that one of the findings is that the “ethical preference for inaction is primarily concentrated in Western cultures” (target article, sect. X, para. X). However, how stable is this finding?

Interestingly, Awad, Dsouza, Shariff, Rahwan, and Bonnefon (2020), using data collected from the same source as Awad et al. (2018), examine the actual trolley problem rather than



studies inspired by it. They find that the preference for the inaction alternative (e.g., for not switching the trolley to kill one person in order to save five) tends to be greater in Eastern cultures than in Western ones. Moreover, even within these cultures, there is variation, such that some countries in the Eastern culture have a greater preference for inaction than those in the Western culture. Separately they note that acceptability of the action alternative has increased over time. Given the variance of the findings and their instability over time, what can be their relevance if not understood by means of underlying theoretical constructs?

Our argument is that the value of a finding lies in its ability to lead to theoretical understanding. For example, a higher-level theoretical explanation arising from the effects observed in moral dilemma experiments might center on a theoretical construct such as norms of social responsibility. We might hypothesize that this construct explains why the effects occur and why they might not occur in different cultures and contexts with different norms of social responsibility. Any theory is of course subject to revision through evaluation of additional evidence, but the essence of theory lies in the development of explanatory constructs that are not tied to any specific set of variables (Calder et al., 2021). And it is theory, not previously observed variable relationships, that should guide application (Calder, Phillips, & Tybout, 1981; Gal & Rucker, 2022, 2023) and that can reconcile seemingly disparate findings.

Consider, for example, the attraction effect (Huber, Payne, & Puto, 1982), where adding a choice alternative (the decoy), which is similar to but inferior to one of the alternatives (the target) but not the other (the competitor), increases the choice of the similar but superior alternative. This effect occurs reliably when the features of the choice alternatives are presented numerically but not when one of the features is presented perceptually. These findings indicate when the effect occurs, but not why numerical and perceptual information produce different effects. As a result, interest in this paradigm waned.

Introducing a theoretical framework that casts the task employed to demonstrate the attraction effect as involving cognitive resource allocation addresses this issue. The hypothesis is that choice is disambiguated by adopting an effort conservation goal that is accessible for numerical but not the more-complex perceptual information provided a means of documenting a perceptual attraction effect. It entailed reducing the effort required to make the decoy accessible as a comparison standard for the target and thus adoptable in making a choice (He & Sternthal, 2023). This framework has also been shown to account for why repeating a single persuasive message has a different effect than repeating different statements that are either truthful or false, and why the depletion effect and its reversal occur (Calder, He, & Sternthal, 2023). The introduction of a theoretical framework thus results in cumulative knowledge across seemingly disparate effects and paradigms.

We contend that confidence in the explanatory power and scope of theory is critical to reconciling seemingly disparate findings and to application in the social and behavioral sciences. Confidence in research findings arises through theoretical understanding not from attempting to map variable to variable relationships.

**Competing interest.** None.

## References

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64.

- Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences of the United States of America*, 117(5), 2332–2337.
- Calder, B. J., Brendl, C. M., Tybout, A. M., & Sternthal, B. (2021). Distinguishing constructs from variables in designing research. *Journal of Consumer Psychology*, 31(1), 188–208.
- Calder, B. J., He, S., & Sternthal, B. (2023). Using theoretical frameworks in behavioral research. *Journal of Business Research*, forthcoming.
- Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, 8(2), 197–207.
- Gal, D., & Rucker, D. D. (2022). Experimental validation bias limits the scope and ambition of applied behavioural science. *Nature Reviews Psychology*, 1(1), 5–6.
- Gal, D., & Rucker, D. D. (2023). Behavioral winter: Disillusionment with applied behavioral science and a path to spring forward. *Behavioral and Brain Sciences*, forthcoming.
- He, S., & Sternthal, B. (2023). Beyond numbers: An ambiguity–accessibility–applicability–accessibility framework to explain the attraction effect. *Journal of Consumer Research*, forthcoming.
- Huber, J., Payne, J. W., & Puto, C. P. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Marketing Research*, 9, 90–98.

## The future of experimental design: Integrative, but is the sample diverse enough?

Sakshi Ghai<sup>a</sup> and Sanchayan Banerjee<sup>b\*</sup>

<sup>a</sup>Department of Psychology, Cambridge University, Cambridge, UK and

<sup>b</sup>Environmental Economics, Institute for Environmental Studies, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

sg912@cam.ac.uk; <https://www.psychol.cam.ac.uk/staff/sakshi-ghai>

S.Banerjee@vu.nl; <https://research.vu.nl/en/persons/sanchayan-banerjee>

\*Corresponding author.

doi:10.1017/S0140525X23002212, e42

### Abstract

Almaatouq et al. propose an “integrative approach” to increase the generalisability and commensurability of experiments. Yet their metascientific approach has one glaring omission (and misinterpretation of) – the role of sample diversity in generalisability. In this commentary, we challenge false notions of presumed duality between contexts, population, and diversity, and propose modifications to their design space to accommodate sample diversity.

Almaatouq et al. propose an “integrative approach” to increase the generalisability and commensurability of experiments. They suggest systematically sampling and testing a subset of experiments – chosen randomly from a design space – to deduce inferences about the population of all potential experiments. Yet their metascientific approach, which loosely translates the “potential outcomes framework” underlying experiments (Holland, 1986; Neyman, 1923; Rubin, 1977) to the science of experimentation itself, has one glaring omission (and misinterpretation of) – the role of sample diversity in reasoning generalisability. Their suggestion that “an explicit, systematic mapping of research designs to points in the design space (research cartography) ensures commensurability” (target article, sect. 3.1, para. 6) is, at best, an

ex-post exercise to ensure validity, based on the myopic assumption that “there is nothing special about the subjects...in principle, what goes for subjects also holds for contexts” (target article, sect. 2.2, para. 4). In this commentary, we challenge these false notions of subsumed duality between contexts, population, and sample diversity, and highlight the importance of diversity in ensuring commensurability. We propose modifications to their design space to incorporate sample diversity.

Almaatouq et al. outline their design space as a Cartesian product of two factors – *population* which is “a set of measurable attributes that characterizes the sample of participants,” and *context* which is a “set of independent variables hypothesized to moderate the effect in question as well as the nuisance parameter” (target article, sect. 2.2, para. 2). Their conceptualisation, however, fails to account for myriad factors arising from the sample and sampling technique itself, which affects the scope of any experiment. We outline three challenges resulting from this.

First, as defined, “population” lacks accountability of representativeness, such as cultural outliers in social and behavioural science experiments, a point that has been argued extensively by critics to metatheories of behaviours (Arnett, 2009; Henrich, Heine, & Norenzayan, 2010). This evident oversight on diversity undermines the role that sample features can play in introducing biases to experiments, invariably leading to methodological narrowness, generating spurious and misleading results (Gurven, 2018; Rad, Martingano, & Ginges, 2018). An integrative framework, therefore, must measure diversity, both between- and within-countries (Ghai, Fassi, Awadh, & Orben, 2023). Without a careful consideration of representativeness in selected sampling approaches and the match of samples to population – be it through crowdsourcing platforms or distributed collaborative networks of different laboratories or sophisticated machine-learning algorithms – integrative experimental techniques will continue to yield noisy and biased results, inapplicable beyond specific population samples.

Second, integrative techniques must grapple with the limitations of not only imperfect sampling approaches but also the limiting assumptions in current disciplinary theories (Medin, Ojalehto, Marin, & Bang, 2017). This is particularly important in the context of Almaatouq et al., who cite that the “ultimate goal” of experiments is to arrive at a comprehensive theoretical understanding of experimental insights. Nonetheless, here, the authors assume (falsely) that metatheories emerging from the design space will naturally lead to heterogeneity and guarantee commensurability. While mapping theoretical boundaries and engaging in meta-metatheoretical reflections, in applying integrative experimental approaches, can be valuable for understanding the generalisability of existing theories, this does not necessarily address underlying structural issues contributing to the lack of theoretical diversity (Haefel & Cobb, 2022). Our critique speaks broadly to need for behavioural sciences to see and reason complex adaptive systems with diverse samples (see Banerjee & Mitra, 2023; Hallsworth, 2023).

Third, acknowledging diversity is important since the costs of running integrative metaexperiments are largely unequal, thereby excluding researchers and relevant stakeholders in the Global South from generating integrated experimental insights. Since, Almaatouq et al. suggest that an “integrative approach would start

by identifying the dimensions... as suggested ... by prior research” (target article, sect. 3.1, para. 4), it is likely this space will then suffer from publication biases. Their claim that an integrative approach “will actually broaden the range of people involved in behavioral research” (target article, sect. 5.7, para. 1) is, at best, misguided, given this drawback. Integrative methods may be transferable but such initiatives are expected to be concentrated and accepted in Western contexts mostly (Singh, 2022). Thus, while we share the optimism for large-scale collaborative science, we are less confident on the ability to draw robust, generalisable conclusions by relying on integrative approaches only. Overcoming the epistemic marginalisation of underrepresented groups in integrative experimental designs arguably is important to achieve this.

In view of these critiques, we propose a modification to their design space that we think is necessary to unlock the power of the integrative approach.

Our proposition relates to explicitly measuring (sample) diversity to quantify the heterogeneity within the sample, advancing the goals of Almaatouq et al. One approach might be using a scalar measure of representativeness for all population and contextual characteristics, for any given experimental point. This scalar measure can then be used to transform and reweigh the design space for generalisability. For example, the authors’ conceptualisation of the population space merely accounts “for a set of measurable attributes” rather than a *rich and diverse* set of measurable attributes “that characterizes the sample of participants” (target article, sect. 2.2, para. 2). As such, their original design space is unduly influenced by certain population subsamples more than others. Here, sampled experimental points cannot be fully representative of all potential experiments. Nonetheless, our approach of first measuring diversity as a scalar index, to then transform these factors of the design space, just like a weighted sampling approach, increases reliability of integrative experiments (Deffner, Rohrer, & McElreath, 2022). One limitation of this approach is that scalar measures of diversity may vary depending on the population, as what even counts as diverse samples will widely differ between the Global North and South contexts (Ghai, 2022). Here we call on the field to develop new ways to increase global diversity and analyse which of these might work best in optimising the design space (Tang, Suganthan, & Yao, 2006).

Ultimately, given that many new sources of knowledge are likely to emerge from the Global South and that these are likely to deviate from Western-centric behavioural insights (Adetula, Forscher, Basnight-Brown, et al., 2022), accounting for the sample’s diversity will truly enhance the scope of integrative experimental methods. The future of experimental design must not only be integrative but also diverse and inclusive.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Arnett, J. J. (2009). The neglected 95%, a challenge to psychology’s philosophy of science. *American Psychologist*, 64(6), 571–574. <https://doi.org/10.1037/a0016723>
- Banerjee, S., & Mitra, S. (2023). Behavioural public policies for the social brain. *Behavioural Public Policy*, 1–23.
- Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, 5(3). <https://doi.org/10.1177/25152459221106366>

- Ghai, S. (2022). Expand diversity definitions beyond their Western perspective. *Nature*, 602(7896), 211. <https://doi.org/10.1038/d41586-022-00330-0>
- Ghai, S., Fassi, L., Awadh, F., & Orben, A. (2023). Lack of sample diversity in research on adolescent depression and social media use: A scoping review and meta-analysis. *Clinical Psychological Science*, 0(0). <https://doi.org/10.1177/21677026221114859>
- Curven, M. D. (2018). Broadening horizons: Sample diversity and socioecological theory are essential to the future of psychological science. *Proceedings of the National Academy of Sciences of the United States of America*, 115(45), 11420–11427.
- Haefel, G. J., & Cobb, W. R. (2022). Tests of generalizability can diversify psychology and improve theories. *Nature Reviews Psychology*, 1(4), 186–187. <https://doi.org/10.1038/s44159-022-00039-x>
- Hallsworth, M. (2023). A manifesto for applying behavioural science. *Nature Human Behaviour*, 7(3), 1–13.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.
- Adetula, A., Forscher, P. S., Basnight-Brown, D., Azouaghe, S., & IJzerman, H. (2022). Psychology should generalize from – not just to – Africa. *Nature Human Behaviour*, 1, 370–371. <https://doi.org/10.1038/s44159-022-00070-y>
- Medin, D., Ojalehto, B., Marin, A., & Bang, M. (2017). Systems of (non-) diversity. *Nature Human Behaviour*, 1(5), 88.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Annals of Agricultural Sciences*, 1–51.
- Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of *Homo sapiens*: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 115(45), 11401–11405.
- Singh, L. (2022). Navigating equity and justice in international collaborations. *Nature Human Behaviour*, 1, 372–373. <https://doi.org/10.1038/s44159-022-00077-5>
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1), 1–26.
- Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, 65, 247–271.

## Individual differences do matter

Stefan Glasauer\* 

Computational Neuroscience, Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany  
[stefan.glasauer@b-tu.de](mailto:stefan.glasauer@b-tu.de)  
<https://www.b-tu.de/en/computational-neuroscience>

\*Corresponding author.

doi:10.1017/S0140525X2300242X, e34

### Abstract

The integrative experiment design proposal currently only relates to group results, but downplays individual differences between participants, which may nevertheless be substantial enough to constitute a relevant dimension in the design space. Excluding the individual participant in the integrative design will not solve all problems mentioned in the target article, because averaging results may obscure the underlying mechanisms.

Many of us probably have experienced that fear responses to certain stimuli, such as spiders or snakes, are not the same for everyone, even when we are in the same situation. The difference between such individual responses is evidently not negligible compared to those elicited by situation-specific conditions. However, this is what Almaatouq et al. suggest when they

downplay the role of individual differences in their otherwise excellent proposal of integrative experiment design (target article, sect. 2.2, para. 4). In psychology, the so-called person–situation debate has been going on for more than a century, and Almaatouq et al. obviously belong to the “situation” faction. Although influential studies such as the famous work of Mischel (1968) have given the “person” faction a backlash since the second half of the twentieth century, there has been a recent resurgence in personality psychology that emphasizes the importance of individual differences (see Roberts & Yoon, 2022). Also in other fields, the importance of the individual has been recognized for quite some time. Recent approaches towards personalized medicine and therapy are probably the most prominent example. A PubMed search for these keywords in title or abstract yields over 33,000 results within the last 20 years. Personalized approaches can be found, for example, in treatment of certain forms of leukemia (Bazinet & Kantarjian, 2023) or in autoimmune diseases (Miner & Fitzgerald, 2023). In other fields, the importance of the individuality of behavior has been recognized as well: Even *Drosophila* shows idiosyncratic behavioral differences, which are caused by developmental variations in brain wiring rather than genetic factors (Linneweber et al., 2020).

However, it is not necessary to resort to personality to see that individual differences found experimentally in a single experiment can be as large as those possibly found by changing the situation, thus contradicting Brunswick’s assumption cited in the target article. In magnitude reproduction tasks, participants often overestimate small magnitudes and underestimate large ones, a perceptual bias termed *central tendency* (Hollingworth, 1910). The central tendency is quantified by 1-slope of reproduced magnitude plotted over stimulus magnitude. In a duration reproduction experiment (Glasauer & Shi, 2022) with randomized stimuli, individual differences in central tendency ranged from 0.1 to about 0.7 (average 0.44), thus covering almost the whole range from veridical reproduction (central tendency 0) to stimulus independence (central tendency 1). In a different experimental situation, when the temporal order of the trials followed a random walk, the average central tendency decreased to 0.11, which means that in this condition duration reproduction was almost veridical.

This example shows that individual differences within one situation can be as large, or even larger, as differences caused by a change of situation. Thus, individual differences can be large enough to be eligible as separate design dimension. However, targeted sampling along that dimension is hardly possible – which criterion would tell us which individual to test? Moreover, the inclusion of the individual as explicit dimension in the design space may often be not viable, because individuals correspond to discrete variables, and “meaningfully distinguishing between the various settings of a discrete variable could require dozens or even hundreds of descriptors” (Eyke, Koscher, & Jensen, 2021). Notably, this is a problem that might also affect the proposed dimension *population*, which could also be composed of many more different descriptors than can be tested.

The usual way to deal with individual differences is to model them as random effect (Yarkoni, 2022). In this view, individual differences are treated as variation of unknown origin without interest in the question. Thus, the solution to the problem of individual differences is to simply constrain



the theories to the group level so that individual variation eventually averages out when groups are sufficiently heterogeneous and large.

In the study mentioned above (Glasauer & Shi, 2022) instead of treating individual differences as random effect, we could explain them by a Bayesian model which assumes that participants entertain *individual* beliefs about how sequential trial-by-trial stimuli are generated. The same model also predicted the massive change in central tendency from one condition to the other (0.44 down to 0.11) based on the individual differences identified in the first condition, that is, without changing the individual model parameters. Thus, behavioral differences observed in different experimental situations do not necessarily indicate actual changes in participants' characteristics (such as beliefs, or personality). This latter conclusion does not depend on whether our theoretical model is correct: The model demonstrates that it is possible.

Thus, while a group-level theory might explain the situation-dependent change, this approach would not allow for a theory that links observed differences to a variation in *individual* characteristics. The same holds for *Drosophila* behavior: Considering only average behavior could not reveal that individual differences are not just random but have the distinct reason of being caused by variation in neuronal wiring.

A possible solution for including individual differences as important information in the integrative experiment design proposal could be to consider the interindividual variability of the variable of interest as additional input for the sampling procedure. Large variability could on the one hand indicate that the particular context of an experiment is not sufficiently constrained, thus leaving too much space for individual differences, indicating design dimensions that have not been included. On the other hand, in the example above the point in design space that resulted in small variability was hiding possible interindividual differences, and thus, from the perspective of theory building, the point with large variability might be the more interesting one. Thus, a sampling procedure that considers interindividual variability could help in defining regions in design space that provide either situations with homogeneous behavioral results, or situations unraveling differences that are of interest for any theory interested in the individual.

**Financial support.** This study was funded in part by Deutsche Forschungsgemeinschaft (GL 342/3-2).

**Competing interest.** None.

## References

- Bazin, A., & Kantarjian, H. M. (2023). Moving toward individualized target-based therapies in acute myeloid leukemia. *Annals of Oncology*, 34(2), 141–151. doi:10.1016/j.annonc.2022.11.004
- Eyke, N. S., Koscher, B. A., & Jensen, K. F. (2021). Toward machine learning-enhanced high-throughput experimentation. *Trends in Chemistry*, 3(2), 120–132. doi:10.1016/j.trechm.2020.12.001
- Glasauer, S., & Shi, Z. (2022). Individual beliefs about temporal continuity explain variation of perceptual biases. *Scientific Reports*, 12, 10746. doi:10.1038/s41598-022-14939-8
- Hollingworth, H. L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, 7(17), 461–469.
- Linneweber, G. A., Andriatsilavo, M., Dutta, S. B., Bengochea, M., Hellbruegge, L., Liu, G., ... Hassan, B. A. (2020). A neurodevelopmental origin of behavioral individuality in the *Drosophila* visual system. *Science (New York, N.Y.)*, 367(6482), 1112–1119. doi:10.1126/science.aaw7182



Miner, J. J., & Fitzgerald, K. A. (2023). A path towards personalized medicine for auto-inflammatory and related diseases. *Nature Reviews Rheumatology*, 19(3), 182–189. doi:10.1038/s41584-022-00904-2

Mischel, W. (1968). *Personality and assessment*. Wiley.

Roberts, B. W., & Yoon, J. Y. (2022). Personality psychology. *Annual Review of Psychology*, 73, 489–516. doi:10.1146/annurev-psych-020821-114927

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1: 1–78. doi: 10.1017/ S0140525X20001685.

## Getting lost in an infinite design space is no solution

Mario Gollwitzer\*  and Johannes Prager 

Department of Psychology, Ludwig-Maximilians-Universität München, München, Germany

mario.gollwitzer@psy.lmu.de

jo.prager@psy.lmu.de

[https://www.psy.lmu.de/soz\\_en/team/professors/mario-gollwitzer/index.html](https://www.psy.lmu.de/soz_en/team/professors/mario-gollwitzer/index.html)

[https://www.psy.lmu.de/soz\\_en/team/academic-staff/prager/index.html](https://www.psy.lmu.de/soz_en/team/academic-staff/prager/index.html)

\*Corresponding author.

doi:10.1017/S0140525X23002236, e44

### Abstract

Almaatouq et al. argue that an “integrative experiment design” approach can help generating cumulative empirical and theoretical knowledge. Here, we discuss the novelty of their approach and scrutinize its promises and pitfalls. We argue that setting up a “design space” may turn out to be theoretically uninformative, inefficient, and even impossible. Designing truly diagnostic experiments provides a better alternative.

Almaatouq et al. argue that research findings in the behavioral and social sciences are rarely conclusive and even less integrative or cumulative. They claim that these insufficiencies largely originate from a flawed approach to research design and the “one-at-a-time” procedure of planning and conducting experiments. Almaatouq et al. propose an alternative: “Integrative experimentation.”

While we share many of the concerns raised in the target article, we have reservations about the proposed alternative. First, the “integrative experimentation” idea is – in principal – not novel. Similar demands for integrative, diagnostic, and cumulative experimentation have been voiced repeatedly in the past (e.g., Brunswik, 1955; Campbell, 1957; Meehl, 1978; Platt, 1964). That said, truly cumulative research endeavors and rigid theory testing has not really been a strength of psychological science so far. Walter Mischel once called this the “toothbrush problem” of psychological science: “Psychologists treat other peoples’ theories like toothbrushes – no self-respecting person wants to use anyone else’s” (Mischel, 2008). Not much has changed since then – at least not until the replication crisis hit psychology full force.

Second, and more importantly, constructing the “design space,” the first step in Almaatouq et al.’s integrative experimentation approach, is practically very difficult, if not impossible. Where should the dimensions that constitute a design space

come from? One problem is that these dimensions are either theoretical constructs themselves or at least defined and informed by theoretical assumptions, and that their number is potentially infinite. Consider the group synergy/group performance example discussed in the target article: Design space dimensions such as “social perceptiveness” or “cognitive-style diversity” are theoretical constructs. The theories that underlie these concepts provide definitions, locate them in a nomological network, and allow the construction of measurement tools. The latter is particularly important because a measurement theory is necessary to test the construct validity of operationalizations (measures and/or manipulations) for a given construct (e.g., Fiedler, McCaughey, & Prager, 2021). So, each design-space dimension turns out to have its own design space. Constructing design spaces may therefore lead to infinite regresses.

Third, and related to our previous point, setting up a design space is necessarily contingent on both the focal hypotheses *as well as* all accompanying auxiliary assumptions, some of which are relevant for testing (and building) a theory, while others are not. For instance, operationalizing “cognitive-style diversity” either as the sum of the within-team standard deviation across cognitive styles (Aggarwal & Woolley, 2013) or as the average team members’ intrapersonal cognitive-style diversity score (Bunderson & Sutcliffe, 2002) may be relevant for testing a measurement theory, yet irrelevant for testing a substantive theory. Almaatouq et al.’s claim that “All sources of measurable experimental-design variation are potentially relevant, and decisions about which parameters are relatively more or less important are to be answered empirically” (target article, sect. 3, para. 1) disregards the (in our view, important) distinction between conceptually relevant and conceptually irrelevant design space dimensions (see Gollwitzer & Schwabe, 2022).

Fourth, all design-space decisions necessarily require a universal metatheory. Such a metatheory does not exist, and it is unlikely to appear in the future. To be clear: We do not oppose to the idea of integrative experimentation. Indeed, we second Almaatouq et al.’s (and Walter Mischel’s) claim for more theory building and integration (“cumulative science”). But without a metatheory that can help us define the design space, a theoretically guided and targeted “one-at-a-time approach” may eventually yield more solid cumulative evidence than a theory-free “integrative experimentation” approach.

Fifth, and finally, since resources and the number of participants required for conducting experiments are limited, randomly combining potentially infinite design factors is not efficient. Atheoretical sampling from the design space requires a tremendous amount of resources, which are better spent testing theories in a most rigid and truly falsification-oriented fashion.

Instead of randomly sampling from an infinite design space, a more useful rationale for setting up an integrative experimental series requires acknowledging a hierarchy regarding the informativeness and discriminability of potential design factor decisions (Fiedler, 2017). Design factors that can – either logically or theoretically – be defined as more crucial than others (e.g., because they relate to a core assumption of a theory or provide a strong test of the boundary conditions of a theory) should be prioritized over those that are more marginal. For instance, if a theory predicted that interacting groups outperform nominal groups more strongly in conjunctive than in additive tasks, an experimental

series might focus on varying conjunctive and additive tasks (or varying different forms of group interaction), while holding other features, which are less relevant to a rigid test of this hypothesis, constant (e.g., culture) or treating these features as random factors (e.g., sample characteristics).

Indeed, modern learning methodology can support the identification of the most informative and diagnostic design factors. However, such models do not learn from random and independent data, but from results depending on design factors that were a priori defined, scaled, and considered relevant (or irrelevant) as well as design-space boundaries set by the experimenters. To conclude, conducting conclusive research means (1) improving the precision of formulated hypotheses, (2) specifying metahypotheses from which (at least some) auxiliary assumptions and boundary conditions can be deduced, and (3) pursuing a systematic, exclusive, and diagnostic testing strategy. From our perspective, systematic research design does not mean searching through a maximally affordable design space, but carefully designing experiments that *exclude* explanations, alternative phenomena, or assumptions. This goal can rather be achieved by systematic exclusion than exploration of possibilities (Wason, 1960). Scaling up such diagnostic experimental series would not require changing the methodological paradigm: On the contrary, truly integrative, metatheoretical, conclusive, and cumulative research is nothing less than proper execution of a long-known and widely accepted experimental methodology.

**Financial support.** This work was supported by a grant provided by the Deutsche Forschungsgemeinschaft (SPP 2317 – project No. 467852570) to the first author.

**Competing interest.** None.

## References

- Aggarwal, I., & Woolley, A. W. (2013). Do you see what I see? The effect of members’ cognitive styles on team processes and errors in task execution. *Organizational Behavior and Human Decision Processes*, 122(1), 92–99. <https://doi.org/10.1016/j.obhdp.2013.04.003>
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217. <https://doi.org/10.1037/h0047470>
- Bunderson, J. S., & Sutcliffe, K. M. (2002). Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects. *Academy of Management Journal*, 45(5), 875–893. <https://doi.org/10.2307/3069319>
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. <https://doi.org/10.1037/h0040950>
- Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 12(1), 46–61. <https://doi.org/10.1177/1745691616654458>
- 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*, 16(4), 816–826. <https://doi.org/10.1177/1745691620970602>
- Gollwitzer, M., & Schwabe, J. (2022). Context dependency as a predictor of replicability. *Review of General Psychology*, 26(2), 241–249. <https://doi.org/10.1177/10892680211015635>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Mischel, W. (2008). The toothbrush problem. *APS Observer* [Online Resource]. <https://www.psychologicalscience.org/observer/the-toothbrush-problem>
- Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science (New York, N.Y.)*, 146(3642), 347–353. <https://doi.org/10.1126/science.146.3642.347>
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, 12, 129–140. <https://doi.org/10.1080/17470216008416717>

# Neuroadaptive Bayesian optimisation can allow integrative design spaces at the individual level in the social and behavioural sciences... and beyond

Rianne Haartsen\* , Anna Gui   
and Emily J. H. Jones 

Centre for Brain and Cognitive Development, Birkbeck, University of London, London, UK

[r.haartsen@bbk.ac.uk](mailto:r.haartsen@bbk.ac.uk)

[agui01@mail.bbk.ac.uk](mailto:agui01@mail.bbk.ac.uk)

[e.jones@bbk.ac.uk](mailto:e.jones@bbk.ac.uk)

<https://cbcd.bbk.ac.uk/people/scientificstaff/rienne-haartsen>

<https://cbcd.bbk.ac.uk/people/scientificstaff/anna-gui>

<https://sites.google.com/view/bondcbcd>

<https://sites.google.com/view/bonds-project/home>

\*Corresponding author.

doi:10.1017/S0140525X23002388, e45

## Abstract

Almaatouq et al. propose an integrative experiment design space combined with large samples for scientific advancement. We argue recent innovative designs combining closed-loop experiment designs and Bayesian optimisation allow for integrative experiments at an individual level during a single session, circumventing the necessity for large samples. This method can be applied across disciplines, including developmental and clinical research.

Almaatouq et al. propose that to improve the generalisability and efficiency of the research in the social and behavioural sciences, experiments should be systematically selected from a large design space. The authors argue that this approach requires the application of the paradigm to large samples and thus extensive collaboration. However, we believe that large samples are useful but not necessary to systematically probe the topography of experimental space. Recently, a conceptually comparable approach called neuroadaptive Bayesian optimisation (NBO) has been developed in neuroimaging. NBO does not require large samples and can operate on an individual level in real time. We propose this approach can also be generalised to the social and cognitive sciences to further the design proposed by Almaatouq et al.

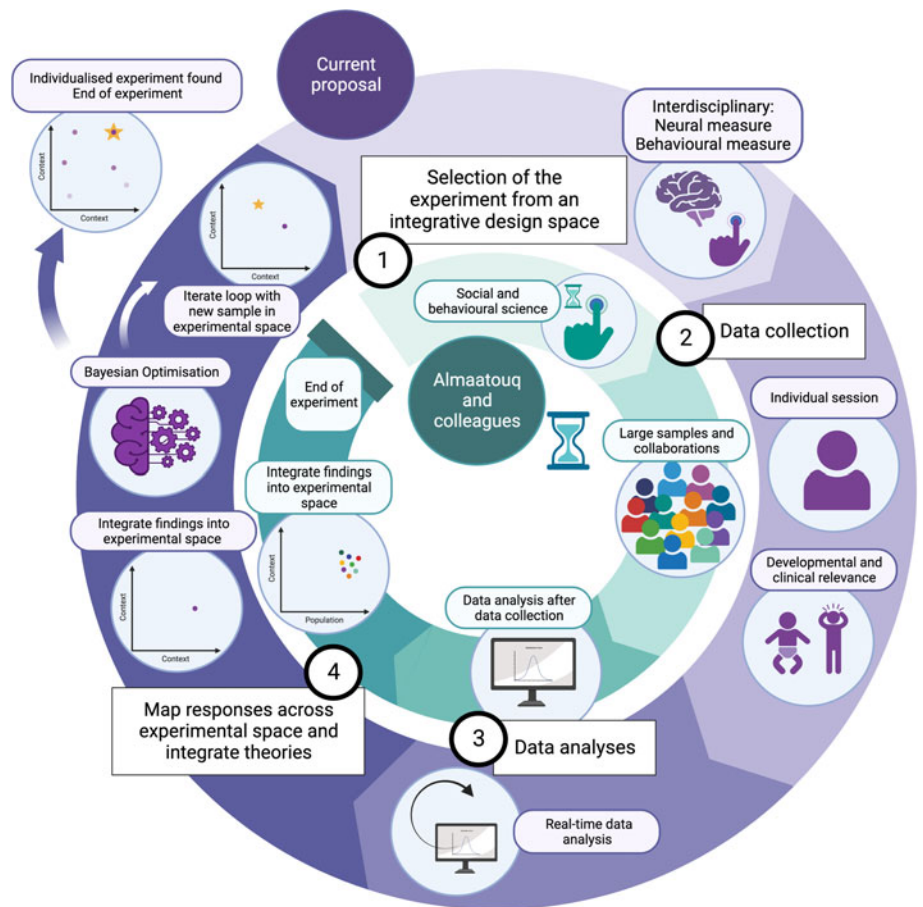
NBO combines real-time analysis of participants' responses and machine learning into a closed-loop design to find the experimental parameters that maximise the target (cognitive or brain) measure while sampling across a large design space (Lorenz, Hampshire, & Leech, 2017) (Fig. 1). Bayesian optimisation (BO) is an active sampling technique that learns from input data collected during a single-experimental testing session. The algorithm uses an acquisition function that balances exploration and exploitation to select an experiment (e.g., a particular combination of characteristics of the presented stimulus) for presentation to the participant from the experiment design space (Brochu, Cora, & de Freitas, 2010). Target measures are collected and modelled

across the design space using Gaussian process regression. Through an iterative process of sampling the search space, a surrogate model of the target response across the full design space is estimated and updated with every sampling. BO will select the next experiment in the design space in an explorative manner early during the session (areas that have not been sampled yet and for which the uncertainty regarding the relationship with the target measure is high) and an exploitative manner towards the end of the session (areas that show the maximum brain/cognitive response and have been sampled previously such that predictability of the response is high). The NBO converges on the area of the design space that maximises the target measure for that individual participant within a few iterations, if the target measure is reliable and the effect size is sufficient. In one early study, researchers constructed an experimental space based on a meta-analysis of existing literature (as suggested by Almaatouq et al.), and used NBO with real-time fMRI analysis to identify cognitive tasks that maximally dissociated between frontoparietal attention networks at the individual level (Lorenz et al., 2018). Thus, NBO brings an integrative experiment design approach to the level of the individual through incorporating the tools of active learning with real-time data analysis.

We argue that integrative experiment designs as proposed may have parallels across multiple fields, and interdisciplinary exchange may be fruitful. NBO is currently being applied in developmental science to examine social development in infants (Gui et al., 2022; Wass & Jones, 2023). Developmental researchers typically preselect a limited range of experimental conditions or stimuli based on specific theories. Similar to issues in social science identified by Almaatouq et al., this has limited progress because each study only probes selected questions and the relationship between theoretical models or different experiment designs remains uncharacterised. In contrast, integrative experiment designs allow experimenters to map brain or behavioural responses across a larger experimental space, including providing out-of-sample predictions for stimuli that are unsampled. This enables developmental researchers to simultaneously test multiple developmental theories through considering their predictions for the variation in responses across a larger space, in accordance with Almaatouq et al.'s proposal. This approach can be extended to other multidimensional spaces with different concepts mapped across dimensions of the search space, for example to explore the influence of contingency or sensitivity on infant attention.

Beyond testing theories in basic science, we think integrative experiment designs may also have clinical utility (Lorenz et al., 2017, 2021). We currently lack objective biomarkers with clinical utility for psychiatric conditions (Loth, 2023). As in the social sciences described by Almaatouq et al., the field of psychiatry is challenged by reproducibility issues that stem from heterogeneity in participant populations, selection of single tasks based on particular theories, and broad analytic flexibility of the resulting data. This leads to difficulty with integrating clinical findings from different studies and hinders biomarker discovery. Integrative experiment designs allow researchers to expand the search space across multiple different tasks or analysis pipelines (Lorenz et al., 2017). Adaptive designs can then be used to select the task and/or pipeline that shows greater individual deviation from population norms for a particular person, similar to Almaatouq et al.'s proposal of mapping individual-level traits across a design space. For example, NBO has recently been used to identify tasks





**Figure 1** (Haartsen et al.). Overview of the integrative experiment design proposed by Almaatouq et al. (inner circle) and how it is operationalised in the neuroadaptive Bayesian optimisation (NBO) approach described in the current proposal (outer circle). This figure was created with [Biorender.com](https://biorender.com).

sensitive to residual network function in individual patients with stroke and higher dissimilarity in responses in patients compared to controls (Lorenz et al., 2021). Similarly, an experimental space can be constructed through characterising the similarity space of the output of different neuroimaging analysis pipelines and identify a pipeline with the most experimental sensitive contrasts, illustrating one way in which the “cartographer” can assign numerical coordinates to different locations in space (Dafflon et al., 2022).

In summary, we argue that the approach proposed by Almaatouq et al. has relevance beyond the boundaries of social and behavioural sciences, and can be extended to the individual level by deploying active learning using real-time feedback during the data collection session itself (Fig. 1). Current implementations in developmental and clinical samples indicate that NBO is particularly promising in samples characterised by high heterogeneity. Further, the NBO approach can be generalised to the social sciences through use of real-time behavioural data collection (i.e., decisions, reaction times, or opinions). Integrative experiment approaches applied at the individual level using real-time data collection, such as NBO, will allow us to conduct reliable research that does not depend on large sample sizes and that can be applied in screening and clinical programmes, answering to Almaatouq et al.’s call to produce unbiased scientific findings that apply to particular real-world contexts.

**Financial support.** This work is supported by EU SAPIENS (grant number 814302), the Economic and Social Research Council (grant number ES/R009368/1), and SFARI GAIINS (grant number 10039678). The results leading to this publication have received funding from the Innovative Medicines

Initiative 2 Joint Undertaking under grant agreement number 777394 for the project AIMS-2-TRIALS. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA and AUTISM SPEAKS, Autistica, SFARI. IMI disclaimer: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. Any views expressed are those of the authors and not necessarily those of the funders.

**Competing interest.** None.





## References

- Brochu, E., Cora, V. M., & de Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. <https://arxiv.org/pdf/1012.2599.pdf>
- Dafflon, J., da Costa, P. F., Vasa, F., Monti, R. P., Bzdok, D., Hellyer, P. J., ... Leech, R. (2022). A guided multiverse study of neuroimaging analyses. *Nature Communications*, 13(1), 3758. <https://doi.org/10.1038/s41467-022-31347-8>
- Gui, A., Throm, E. V., da Costa, P. F., Haartsen, R., Leech, R., & Jones, E. J. H. (2022). Proving and improving the reliability of infant research with neuroadaptive Bayesian optimization. *Infant and Child Development*, 31(5), 1–6. <https://doi.org/10.1002/icd.2323>
- Lorenz, R., Hampshire, A., & Leech, R. (2017). Neuroadaptive Bayesian optimization and hypothesis testing. *Trends in Cognitive Sciences*, 21(3), 155–167. <https://doi.org/10.1016/J.TICS.2017.01.006>
- Lorenz, R., Johal, M., Dick, F., Hampshire, A., Leech, R., & Geranmayeh, F. (2021). A Bayesian optimization approach for rapidly mapping residual network function in stroke. *Brain*, 144(7), 2120–2134. <https://doi.org/10.1093/brain/awab109>
- Lorenz, R., Violante, I. R., Monti, R. P., Montana, G., Hampshire, A., & Leech, R. (2018). Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization. *Nature Communications*, 9(1), 1227. <https://doi.org/10.1038/s41467-018-03657-3>
- Loth, E. (2023). Does the current state of biomarker discovery in autism reflect the limits of reductionism in precision medicine? Suggestions for an integrative approach that

considers dynamic mechanisms between brain, body, and the social environment. *Frontiers in Psychiatry*, 14, 1–12. <https://doi.org/10.3389/fpsy.2023.1085445>

Wass, S., & Jones, E. J. H. (2023). Editorial perspective: Leaving the baby in the bathwater in neurodevelopmental research. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 64(8), 1256–1259. <https://doi.org/10.1111/jcpp.13750>

## Measurement validity and the integrative approach

Wendy C. Higgins<sup>a\*</sup> , Alexander J. Gillett<sup>b</sup> ,  
Eliane Deschrijver<sup>a</sup>  and Robert M. Ross<sup>b</sup> 

<sup>a</sup>School of Psychological Sciences, Macquarie University, Sydney, NSW, Australia and <sup>b</sup>Department of Philosophy, Macquarie University, Sydney, NSW, Australia  
[wendy.higgins@mq.edu.au](mailto:wendy.higgins@mq.edu.au), <https://researchers.mq.edu.au/en/persons/wendy-higgins>

[alexander.gillett@mq.edu.au](mailto:alexander.gillett@mq.edu.au)

[eliane.deschrijver@mq.edu.au](mailto:eliane.deschrijver@mq.edu.au), <https://researchers.mq.edu.au/en/persons/eliane-deschrijver>

[robross46@gmail.com](mailto:robross46@gmail.com), <https://researchers.mq.edu.au/en/persons/robert-ross>

\*Corresponding author.

doi:10.1017/S0140525X23002194, e46

### Abstract

Almaatouq et al. propose a novel integrative approach to experiments. We provide three examples of how unaddressed measurement issues threaten the feasibility of the approach and its promise of promoting commensurability and knowledge integration.

When scientists lack validity evidence for measures, they lack the necessary information to evaluate the overall validity of a study's conclusions.

Flake and Fried (2020, p. 457)

Questionable measurement practices are widespread in the social and behavioural sciences and raise serious questions about the interpretability of numerous studies (Flake & Fried, 2020; Lilienfeld & Strother, 2020; Vazire, Schiavone, & Bottesini, 2022). Because Almaatouq et al. do not explicitly address measurement, we argue that unresolved measurement issues may threaten the feasibility and utility of their integrative approach. Below, we present three measurement concerns.

First, the interpretability of findings from experiments designed using the integrative approach will rely on the use of valid measurements. Consider the “Moral Machine” experiment (Awad et al., 2018, 2020), which Almaatouq et al. describe as “seminal.” Utilising a modified version of the trolley problem, this experiment evaluated participant’s preferences for how autonomous vehicles should weight lives in life-or-death situations based on nine different dimensions. By assessing these dimensions simultaneously and collecting responses from millions of participants, Almaatouq et al. claim that this experiment “offers numerous findings that were neither obvious nor deducible from prior research or traditional experimental designs” (target article, sect. 4.1, para. 2). One of these key findings is that participants are willing to treat people differently based on demographic characteristics when the complexity of a moral decision is increased. However, the validity of this finding

has been questioned because it may be an artefact of the forced choice methodology that was used (Bigman & Gray, 2020). In addition, there is considerable debate in moral psychology about the external validity of the trolley problem and other sacrificial dilemmas (i.e., it is unclear that responses in these tasks predict real-world decisions or ethical judgements; Bauman, McGraw, Bartels, & Warren, 2014; Bostyn, Sevenhant, & Roets, 2018). Thus, to our minds, this example demonstrates that no matter how large and integrative an experiment might be, evaluating the validity of the measurements is essential.

Second, the construction of design spaces and the mapping of experiments onto them relies on valid measurement of design space dimensions. However, the validity of measurements, including those obtained from widely used measures, cannot be assumed. Consider Almaatouq et al.’s identification of social perceptiveness as a relevant dimension of group synergy research. They cite four studies that measured social perceptiveness using the Reading the Mind in the Eyes Test (RMET; Almaatouq, Alsobay, Yin, & Watts, 2021; Engel, Woolley, Jing, Chabris, & Malone, 2014; Kim et al., 2017; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010). However, it is unclear what psychological constructs the RMET measures. While the RMET has been used to measure multiple dimensions of social cognition, including “theory of mind,” “emotion recognition,” “empathy,” “emotional intelligence,” “mindreading,” “mentalising,” and “social perceptiveness,” there is ongoing debate about the relationship between these constructs and which, if any, of them the RMET actually measures (Kittel, Olderbak, & Wilhelm, 2022; Oakley, Brewer, Bird, & Catmur, 2016; Silverman, 2022). Moreover, despite the extensive use of the RMET (cited over 7,000 times according to Google Scholar), serious questions have been raised about the reliability and validity of RMET scores (Higgins, Ross, Langdon, & Polito, 2023; Higgins, Ross, Polito, & Kaplan, 2023; Kittel et al., 2022; Olderbak et al., 2015). This means that any integrative experiment that uses the RMET to measure social perceptiveness as a dimension of group synergy research will be very difficult to interpret. Given that vast swathes of measures used in psychological and social science research lack good validity evidence (Flake & Fried, 2020), analogous validity concerns are likely to exist for measures of many dimensions of a given design space. Thus, measurement validation is a critical and nontrivial consideration for the construction and implementation of the design spaces at the heart of the integrative approach. Moreover, given that design spaces are likely to include large numbers of dimensions, a coherent strategy to handle these issues must be developed otherwise the integrative approach risks becoming unmanageable in terms of magnitude and complexity.

Third, measurement incommensurability poses a substantial challenge to the feasibility and utility of the integrative approach because knowledge integration relies on valid *and* commensurable measurements. Consider depression, one of the most prevalent mental health conditions worldwide (Herrman et al., 2019). Fried, Flake, and Robinaugh (2022) recently identified over 280 different depression measures. Extensive variability in the symptoms assessed by these measures forced them to conclude that different depression measures “seem to measure different ‘depressions’” (p. 360). Moreover, they found that depression measures frequently fail to show measurement invariance, meaning that they might measure different things when used in different groups or contexts. Fried and colleagues’ examination of depression measures is an unusually thorough demonstration of just how serious measurement incommensurability problems

can be. Nonetheless, there are indications that validity and commensurability problems extend to a diverse range of research areas which, troublingly, are also pertinent to human welfare, including child and adolescent psychopathology (Stevanovic et al., 2017); race-related attitudes, beliefs, and motivations (Hester, Axt, Siemers, & Hehman, 2023); and well-being (Alexandrova & Haybron, 2016). While Almaatouq et al. claim that their integrative approach “intrinsically promotes commensurability and continuous integration of knowledge” (target article, abstract), it is unclear how the approach can feasibly address incommensurability arising from the use of disparate measures and violations of measurement invariance. Left unaddressed, measurement incommensurability might substantially curtail the knowledge integration potential of the proposed approach.

To summarise, although we are sympathetic to Almaatouq et al.’s ambitious attempt to tackle the substantial challenges in the psychological and behavioural sciences, their lack of engagement with the measurement literature raises serious questions about their approach. If it is to deliver its intended benefits of increased commensurability and knowledge integration, then measurement must be addressed explicitly. It is unclear to us whether this can be achieved while maintaining the feasibility of the proposed integrative approach.

**Financial support.** This work was supported by an Australian Government Research Training Program (RTP) Scholarship (W. C. H.), a Macquarie University Research Excellence Scholarship (W. C. H.), a Discovery Early Researcher Award (DECRA) by The Australian Research Council (ARC) (E. D., grant number DE220100087), and the John Templeton Foundation (R. M. R., grant number 62631; A.G., grant number 61924).


**Competing interest.** None.

## References

- Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science*, 83(5), 1098–1109. <https://doi.org/10.1086/687941>
- Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences of the United States of America*, 118(36), Article e21101062118. <https://doi.org/10.1073/pnas.2101062118>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2020). Reply to: Life and death decisions of autonomous vehicles. *Nature*, 579(7797), E3–E5. <https://doi.org/10.1038/s41586-020-1988-3>
- Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9), 536–554. <https://doi.org/10.1111/spc3.12131>
- Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, 579(7797), E1–E2. <https://doi.org/10.1038/s41586-020-1987-4>
- Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, 29(7), 1084–1093. <https://doi.org/10.1177/0956797617752640>
- Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS ONE*, 9(12), e115212. <https://doi.org/10.1371/journal.pone.0115212>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
- Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, 1(6), 358–368. <https://doi.org/10.1038/s44159-022-00050-2>
- Herrman, H., Kieling, C., McGorry, P., Horton, R., Sargent, J., & Patel, V. (2019). Reducing the global burden of depression: A Lancet–World Psychiatric Association Commission. *The Lancet*, 393(10189), e42–e43. [https://doi.org/10.1016/S0140-6736\(18\)32408-5](https://doi.org/10.1016/S0140-6736(18)32408-5)
- Hester, N., Axt, J. R., Siemers, N., & Hehman, E. (2023). Evaluating validity properties of 25 race-related scales. *Behavior Research Methods*, 55(4), 1758–1777. <https://doi.org/10.3758/s13428-022-01873-w>

- Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2023). The “reading the mind in the eyes” test shows poor psychometric properties in a large, demographically representative U.S. sample. *Assessment*, 30(6), 1777–1789. <https://doi.org/10.1177/10731911221124342>
- Higgins, W. C., Ross, R. M., Polito, V., & Kaplan, D. M. (2023). Three threats to the validity of the reading the mind in the eyes test: A commentary on Pavlova and Sokolov (2022). *Neuroscience and Biobehavioral Reviews*, 147, 105088. <https://doi.org/10.1016/j.neubiorev.2023.105088>
- Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (pp. 2316–2329). Association for Computing Machinery, Portland, Oregon, USA. <https://doi.org/10.1145/2998181.2998185>
- Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the mind’s eye: A meta-analytic investigation of the nomological network and internal consistency of the “reading the mind in the eyes” test. *Assessment*, 29(5), 872–895. <https://doi.org/10.1177/1073191121996469>
- Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology*, 61(4), 281–288. <https://doi.org/10.1037/cap0000236>
- Oakley, B. F., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the reading the mind in the eyes test. *Journal of Abnormal Psychology*, 125(6), 818–823. <https://doi.org/10.1037/abn0000182>
- Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brennehan, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, 6, 1503. <https://doi.org/10.3389/fpsyg.2015.01503>
- Silverman, C. (2022). How to read “reading the mind in the eyes”. *Notes and Records of the Royal Society of London*, 76(4), 683–697. <https://doi.org/10.1098/rsnr.2021.0058>
- Stevanovic, D., Jafari, P., Knez, R., Franic, T., Atilola, O., Davidovic, N., ... Lakić, A. (2017). Can we really use available scales for child and adolescent psychopathology across cultures? A systematic review of cross-cultural measurement invariance data. *Transcultural Psychiatry*, 54(1), 125–152. <https://doi.org/10.1177/1363461516689215>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2), 162–168. <https://doi.org/10.1177/09637214211067779>
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science (New York, N.Y.)*, 330(6004), 686–688. <https://doi.org/10.1126/science.1193147>

## The social sciences needs more than integrative experimental designs: We need better theories

Moshe Hoffman<sup>a\*</sup> , Tadeq Quillien<sup>b</sup>  
and Bethany Burum<sup>a</sup>

<sup>a</sup>Harvard University, Cambridge, MA, USA and <sup>b</sup>School of Informatics, University of Edinburgh, Edinburgh, UK

Hoffman.moshe@gmail.com, <https://sites.google.com/site/hoffmanmoshe/tadeq.quillien@gmail.com>, <https://sites.google.com/view/tadeq-quillien/homelbethanyburum@gmail.com>, <http://bethanyburum.com>

\*Corresponding author.

doi:10.1017/S0140525X23002297, e47

### Abstract

Almaatouq et al.’s prescription for more integrative experimental designs is welcome but does not address an equally important problem: Lack of adequate theories. We highlight two features theories ought to satisfy: “Well-specified” and “grounded.” We discuss the importance of these features, some positive exemplars, and the complementarity between the target article’s prescriptions and improved theorizing.



We appreciate the target article's criticism of existing social science experimental methods: Results are often presented without clear boundary conditions and without making it easy to compare results across labs. We also appreciate the target article's main prescription for empirically addressing these issues: Systematically exploring the parameters that vary across existing studies and theories. However, in line with recent criticisms of the social sciences (e.g., Muthukrishna & Henrich, 2019), we believe this prescription only takes us halfway; good theorizing is still essential.

We wish to highlight two features of theories that seem especially imperative: (1) Well-specified: Theories should specify a causal process; otherwise it is difficult to generalize out of sample. (2) Grounded: The specified causal process should not "beg for explanation," but instead be explicable in terms of well-understood processes; otherwise it is harder to build up a coherent scientific enterprise and theories might only superficially be adding explanatory power. These two features are intuitively appealing (e.g., Ahn, Kalish, Medin, & Gelman, 1995; Pacer & Lombrozo, 2017), prescribed by philosophers of science (Lakatos, 1978; Pearl, 2000; Woodward, 2003), and can be justified using Bayesian models (Goodman, Ullman, & Tenenbaum, 2011; Griffiths & Tenenbaum, 2009).

Social science theories can satisfy these two properties. Evolutionary game-theoretic approaches to moral psychology provide one exemplar (e.g., Hoffman & Yoeli, 2022; Quillien, 2020). For instance, one account for why we donate to ineffective charities posits that we are partially motivated to give by the reputational benefits, and these reputational benefits can only depend on information that is easy for others to ascertain and agree upon – like whether you gave but not the impact of your gift (Borum, Nowak, & Hoffman, 2020). This account is "grounded" in the sense that it rests on premises that are consistent with known causal processes that do not themselves "beg for explanation" – our morals are subject to evolutionary forces, reputations are a key evolutionary force, and reputations can only depend on information others have and are likely to agree upon (e.g., Boyd, 2018; Cosmides, Guzmán, & Tooby, 2018; DeScioli & Kurzban, 2013; Nowak & Sigmund, 2005). Moreover, this theory is "well-specified" in the sense that it specifies a causal process – reputational benefits shape our moral intuitions via biological or cultural evolution. Finally, this causal process makes clear predictions about generality – for example, we should be *more sensitive* to impact when it comes to our kin or savings decisions (Borum et al., 2020).

Computational models of cognition offer a second exemplar (e.g., Oaksford & Chater, 1994; Quillien & Lucas, 2023; Xu & Tenenbaum, 2007). For instance, in one approach to explaining "anchoring and adjustment" – the fact that numerical estimates can be biased in the direction of an arbitrarily selected value provided one is first asked if the true value is above or below that arbitrarily selected value – anchors are thought to provide a "seed" for a cognitive process that only slowly and effortfully adjusts (Lieder, Griffiths, Huys, & Goodman, 2018a). In this model, people start at the seed, then sample a nearby numerical estimate, check the relative plausibility of this estimate, move toward the new estimate if it is judged to be more plausible, then repeat this process as long as it seems worth the cognitive costs. This explanation is "grounded" in the sense that it rests on plausible assumptions about the scarcity of computational resources and the need to rely on sampling algorithms instead of explicit representations of probability distributions (e.g., MacKay, 2003; Vul, Goodman, Griffiths, & Tenenbaum, 2014). This explanation is "well-specified" in the sense that it specifies

a causal process, which suggests boundary conditions – people are expected to show more of an anchoring bias the fewer computational resources they allocate to the task, say, due to time constraints, cognitive load, or lack of motivation (Lieder, Griffiths, Huys, & Goodman, 2018b).

We note that well-specified theories might already ameliorate the issues motivating the target article. The target article is partially motivated by "one-off studies" that *seem* to contradict each other because they are each run with different parameter settings and conclusions are over-generalized. However, we believe such over-generalizations would be less likely if researchers were forced to limit their conclusions to those warranted by their specified causal process. Consider research on group-synergies: Some studies find individuals work better in isolation, while others find they work better in groups. Such findings may only appear contradictory if we rely on overly broad conclusions – for example, "groups are synergistic." If instead, we focused on causal processes – for example, "groups are useful for division of labor" – and restrict our conclusions to those warranted by the specified causal process – for example, "groups will perform relatively better when the task demands more division of labor" – we would have an easier time reconciling results across labs – for example, because one lab used a task that lent itself more to division of labor.

We also note that the target article's main prescription does not obviate the need for better theorizing. The authors suggest a systematic method of sampling from the parameter values that existing theories predict might matter (perhaps supplemented by "surrogate models" – constructed by training a deep neural network on large amounts of data). However, if existing theories (and surrogate models) are themselves not well-specified or grounded, it is not obvious how the prescribed approach will help us get any closer to theories that are, and without that, it is not obvious that we will not still be missing key latent variables not yet considered. For instance, the target article describes one instance (Agrawal, Peterson, & Griffiths, 2020) where the prescribed approach led to new discoveries in the "Moral Machine" paradigm, such as that people are less likely to save criminals than law-abiding citizens. However, without good theorizing, we are left not knowing what is causing these discoveries, and hence not being able to know their boundary conditions (beyond the dimensions investigated). Nor is it obvious what these discoveries teach us about moral psychology, or the social, cognitive, or biological forces that shape our morals, writ large.

One final note: Without winnowing down the set of theories under consideration, the target article's prescribed method may be unwieldy, since each theory suggests additional variables to systematically investigate. Restricting theories to those that are well-specified and grounded may help reduce the set of theories under consideration, thereby making the prescribed approach more viable.



**Competing interest.** None.

## References

- Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 117(16), 8825–8835.
- Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3), 299–352.
- Boyd, R. (2018). *A different kind of animal: How culture transformed our species*. Princeton University Press.
- Borum, B., Nowak, M. A., & Hoffman, M. (2020). An evolutionary explanation for ineffective altruism. *Nature Human Behaviour*, 4(12), 1245–1257.
- Cosmides, L., Guzmán, R. A., & Tooby, J. (2018). The evolution of moral cognition. In *The Routledge handbook of moral epistemology* (pp. 174–228). Routledge.

- DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, 139(2), 477.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118(1), 110.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116(4), 661.
- Hoffman, M., & Yoeli, E. (2022). *Hidden games: The surprising power of game theory to explain irrational human behavior*. Hachette.
- Lakatos, I. (1978). *The methodology of scientific research programmes*. J. Worrall & G. Currie (Eds.). Cambridge University Press.
- Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018a). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25, 322–349.
- Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2018b). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, 25, 775–784.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3), 221–229.
- Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063), 1291–1298.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4), 608.
- Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, 146(12), 1761.
- Pearl, J. (2000). *Causality*. Cambridge University Press.
- Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, 492, 110204.
- Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*. Advance online publication. <https://doi.org/10.1037/rev0000428>
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245.

## Representative design: A realistic alternative to (systematic) integrative design

Gijs A. Holleman<sup>a\*</sup> , Mandeep K. Dhami<sup>b</sup>,  
Ignace T. C. Hooge<sup>c</sup> and Roy S. Hessels<sup>c</sup> 

<sup>a</sup>Department of Cognitive Neuropsychology, Tilburg University, Tilburg, the Netherlands; <sup>b</sup>Department of Psychology, Middlesex University, London, UK and <sup>c</sup>Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, the Netherlands

[g.a.holleman@tilburguniversity.edu](mailto:g.a.holleman@tilburguniversity.edu)

[m.dhami@mdx.ac.uk](mailto:m.dhami@mdx.ac.uk)

[i.hooge@uu.nl](mailto:i.hooge@uu.nl)

[r.s.hessels@uu.nl](mailto:r.s.hessels@uu.nl); [royhessels@gmail.com](mailto:royhessels@gmail.com)

\*Corresponding author.

doi:10.1017/S0140525X23002200, e48

### Abstract

We disagree with Almaatouq et al. that no realistic alternative exists to the “one-at-a-time” paradigm. Seventy years ago, Egon Brunswik introduced *representative design*, which offers a clear path to commensurability and generality. Almaatouq et al.'s *integrative design* cannot guarantee the external validity and generalizability of results which is sorely needed, while *representative design* tackles the problem head on.

We share Almaatouq et al.'s concerns with the lack of commensurability and generalizability of experimental findings in the social and behavioural sciences. However, we disagree that a “lack of any realistic alternative” (target article, sect. 3, para. 1) existed, which prompted them to propose *integrative design*. Over 70 years ago, Egon Brunswik (1956b, p. 159) saw “intrinsic shortcomings” in “artificial, systematic [experimental] designs” regardless of whether or not these designs were implemented “one-at-a-time” (word in square brackets added). He proposed *representative design* as an alternative. This lays a path towards commensurability and generality as well as a clear vision for theoretically and practically valuable research in psychology.

Brunswik (1944, 1955b, 1956a) questioned the ability of systematic design to yield internally and externally valid results. He argued that variables may be artificially “tied” or “untied,” thus making it impossible to rule out the effect of the confound in the former case and making it impossible to study human functioning in a generalizable way in the latter case. His alternative, representative design retains the “causal texture of the environment” to which the human has adapted and to which the researcher intends to generalize (see Dhami, Hertwig, & Hoffrage, 2004, for a review). For Brunswik, the effect of specific variables should be disentangled at the data analysis rather than data collection stage. By contrast, Almaatouq et al. appear to accept systematic design and only critique its “one-at-a-time” implementation, arguing that results are difficult to compare, aggregate, and generalize. However, their solution to this problem suffers from the same limitations that Brunswik identified with systematic design.

Almaatouq et al.'s notion of the “design space” essentially comprises a large series of environments (combinations of various variables) from countless one-at-a-time experiments. As Brunswik (1955a, 1955b) noted, these will potentially include, at best, environments which are rarely encountered, and most likely, environments that do not (or cannot) exist in the real world. While Almaatouq et al. appear to accept Brunswik's view that the generalizability over situations is equally, if not more, important than that over participants, they fail to recognize the importance of representative stimulus sampling (and construction; see Hammond, 1966). There is no way to know which environments in the design space are representative and which are not. Instead, Almaatouq et al. are preoccupied with reconciling, replicating, or even opening the “file drawer” (target article, sect. 5.2, para. 4) of experimental studies that may lack generality because they were obtained under unrepresentative conditions.

Almaatouq et al. applaud Peterson, Bourgin, Agrawal, Reichman, and Griffiths's (2021) efforts to sample the “space of possible experiments [i.e., gambles] much more densely” (target article, sect. 4.2, para. 2) than before. Yet, they do not question the representativeness of the gambles studied and so the generalizability of the findings remain unknown. Brunswik's representative design (1956a, 1956b) on the contrary, tackles the problem directly; researchers must first define the “reference class” or “universe” of stimuli (tasks/situations, e.g., gambles) about which they want to draw a generalizable conclusion. One then either explicitly samples stimuli from this predefined set or constructs stimuli representative of it. One example where representative design has cast serious doubt over well-established conclusions based on systematic design is given by Juslin, Winman, and Olsson (2000) on the overconfidence phenomenon (for other examples, see Dhami et al., 2004). Representative design can also avoid potential pitfalls of Almaatouq et al.'s method such as the need to configure a

“correct” or “relevant” design space, prioritization of aspects of the space, and keeping the number of possible experiments to a manageable level. Additionally, the use of representative design can be facilitated by virtual reality, and is not hampered by the need for large participants’ pools (since each individual performs multiple trials and data are analysed at the individual level). Simply stated, integrative design cannot guarantee the external validity and generalizability of results which the social and behavioural sciences sorely need, while representative design tackles the problem head on.

To us, the crux of the problem that ails the social and behavioural sciences, which Almaatouq et al. do not address, is: What is the overall goal? Indeed, before any researcher embarks upon designing a study, let alone a paradigm shift in doing research, one ought to consider what their goal is. For Brunswik, the method followed his goal. He envisioned psychology as a science of “organism–environment relationships” (Brunswik, 1943), and he provided the “lens model” framework (Brunswik, 1952, 1955a) for theoretically delineating how individuals are adapted to the environments in which they function (termed *probabilistic functionalism*). Note that we are not calling for an outright rejection of systematic design, but for it to be contextualized within representative design. In our view, psychology need not have one single goal or method, but we do agree with Brunswik that one’s method should follow one’s goal, and that generalizability is important.

Representatively designed experiments can reveal how humans are adapted to their environments. Experiments which then alter specific environmental properties can demonstrate how these adaptation processes are challenged. Thus, representative design requires researchers to delineate environmental properties to understand human environments – something researchers in the social and behavioural sciences rarely do, not even to determine the generality of an existing set of results. Understanding human cognition and behaviour as a function of environmental properties is also highly relevant for practically applicable research, and funding bodies and universities are increasingly rewarding researchers whose findings have impact, thereby providing further incentive for representative design.

In sum, Brunswik was ahead of his time in recognizing that systematic design means that researchers would need to be satisfied by “plausibility generalizations, ... always precarious in nature – or [be] satisfied with results confined to a self-created ivory tower ecology” (1956b, p. 110). He provided a methodological solution to this problem, and a clear theoretical ambition. Unfortunately, his ideas have been largely ignored, forgotten, misunderstood, or even ridiculed (for a history and discussion, see Hammond, 1998; Holleman, Hooge, Kemner, & Hessels, 2020, 2021). By missing the opportunity to build on representative design, Almaatouq et al. themselves contribute to what they see as a fundamental problem in today’s social and behavioural sciences, that is, not “putting things together.”

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50(3), 255.
- Brunswik, E. (1944). Distal focussing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs*, 56(1), 1–49.
- Brunswik, E. (1955a). The conceptual framework of psychology. In R. Carnap, O. Neurath, & C. W. Morris (Eds.), *International encyclopedia of unified science* (Vol. 1, Part 2, pp. 655–760). University of Chicago Press.
- Brunswik, E. (1955b). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193.
- Brunswik, E. (1956a). Historical and thematic relations of psychology to other sciences. *The Scientific Monthly*, 83(3), 151–161.
- Brunswik, E. (1956b). *Perception and the representative design of psychological experiments*. University of California Press.
- Dhmi, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6), 959.
- Hammond, K. R. (1966). Probabilistic functionalism: Egon Brunswik’s integration of the history, theory, and method of psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (Vol. 15, p. 80). Holt, Rinehart and Winston.
- Hammond, K. R. (1998). Ecological validity: Then and now. <https://brunswiksociety.org/wp-content/uploads/2022/06/essay2.pdf>
- Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The “real-world approach” and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, 11, 721.
- Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2021). The reality of “real-life” neuroscience: A commentary on Shamay-Tsoory and Mendelsohn (2019). *Perspectives on Psychological Science*, 16(2), 461–465.
- Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, 107(2), 384.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, 372(6547), 1209–1214.

## Some problems with zooming out as scientific reform

Jessica Hullman\* 

Computer Science, Northwestern University, Evanston, IL, USA  
[jhullman@northwestern.edu](mailto:jhullman@northwestern.edu)

\*Corresponding author.

doi:10.1017/S0140525X23002133, e49

### Abstract

Integrative experimentation will improve on the status quo in empirical behavioral science. However, the results integrative experiments produce will remain conditional on the various assumptions used to produce them. Without a theory of interpretability, it remains unclear how viable it is to address the crud factor without sacrificing explainability.

When faced with social science research, why is it so hard to answer the question: What did we learn from this experiment? A core problem is that many experimenters have come to equate theories with predicting directional associations, which can neither formally ground expectations of when data are surprising nor yield strong experimental tests. Any scientific reform proposal that starts from data generated by sampling a design space and expects to get to good theory misconstrues the role of theory in learning from experiments: To propose a data-generating mechanism with testable implications (Fiedler, 2017; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019).

At the same time, behavioral science is unlikely to change the world if we do not start taking heterogeneity of effects more



seriously (Bryan, Tipton, & Yeager, 2021). Integrative experiment design (target article) elevates heterogeneity by rendering explicitly a larger design space from which experiments are sampled. By applying predictive modeling to test the generalization of surrogate models learned on portions of the space, it addresses the pervasive illusion that models chosen for their explanatory power also predict well (Yarkoni & Westfall, 2017). If adopted, integrative modeling seems well-positioned to improve on the status quo of knowledge generation in many domains.

However, like related proposals that attempt to debias data-driven inferences by “zooming out,” integrative design occupies an in-between territory in which gestures of completeness have conceptual value but struggle to find their footing in the form of stronger guarantees. Here I consider challenges that arise in (1) trying to separate the results sampled from a design space from the assumptions that produce them and (2) trying to achieve a balance between reducing confounds from the crud factor (Meehl, 1990) and drowning in complexity.

### *No such thing as unconditional data*

A presupposition behind integrative experiment design – and related proposals like multiverse analysis, which attempts to amend the limitations of a single analysis by rendering explicitly a design space to sample from (Steege, Tuerlinckx, Gelman, & Vanpaemel, 2016) – is that by zooming out from a narrow focus (on just a few variables, or a single analysis path) and sampling results from a larger space, they will produce unbiased evaluations of a claim. In integrative modeling, tests of surrogate models take the form of prediction problems in a supervised learning paradigm, with the added implied constraint that they must “accurately explain the data researchers have already observed.”

But the theories that arise from integrative experiment design will be conditional on more than just the features used to train them. How to interpret the “tests” of surrogate models is an important degree of freedom, for example. Measures like sample complexity can supply requirements to resolve prediction accuracy within a chosen error bound, but not what bound should constitute sufficient predictive performance, or how it should differ across domains. There is a chicken-and-egg problem in attempting to separate the experimental findings from the definition of the learning problem and sampling approach.

If integrative experiment design also incorporates explanatory methods, and the explanations take the form of causal mechanisms proposed to operate in different regions of the design space, then this explanatory layer may very well make it easier for experimenters to draw on domain knowledge, helping retain predictive accuracy when moving out-of-distribution relative to a “pure prediction” approach. But this is difficult to conclude without defining what makes a surrogate model interpretable.

### *Goldilocks and the crud factor*

Both multiverse analysis and integrative experiment design can seem to presuppose that our prior knowledge can take us just far enough to produce results more complex than current results sections, but not so complicated that we get overwhelmed. The “new kinds of theories” associated with integrative experimentation are meant to “capture the complexity of human behaviors while retaining the interpretability of simpler theories.” This may be possible, but we should be careful not to assume that we can always zoom out until we find the dimensionality that is

considerably greater than the dimensionality of the problem implied by the status quo single experiment, but not so great as to be noncomprehensible to a human interpreter.

If we take seriously Meehl’s notion of the crud factor, we might easily list hundreds of potential influences, for example, on group performance, from interpersonal attractions among group members to their religious orientations to recent current events. Even if the prior literature boils some of these down to encompassing unidimensional summaries (e.g., religious homogeneity) there will be many ways to measure each, and many ways to analyze the results which might have their own consequences. How do we guarantee that the number of choices that matter yield interpretable explanations? To take seriously the promises of approaches like integrative experimentation, we must contextualize them within a theory of interpretability.

Multiverse and integrative experiment design provide solutions that are more relative than precise: Sampling from the larger space better captures our fundamental ontological uncertainty about the true data-generating model than not defining and sampling from the larger space, but cannot eliminate it. By prioritizing data over theory, both approaches gesture toward completeness, but cannot provide guarantees. Under philosophical scrutiny, their clearest value seems to be rhetorical. Consequently the completeness that such methods seem to promise can be misleading.

These points should not discourage adoption of integrative experimentation, which is likely to improve learning from experiments by addressing many important criticisms raised with the status quo. However, as confident but often informal proposals scientific reforms abound, it is always worth deep consideration of what problems are addressed, and what promises, if any, can be made (Devezer, Navarro, Vandekerckhove, & Buzbas, 2021). Integrative experiment design is one way of improving learning from experiments, which can complement but cannot replace the need to clarify what we learn from any experiment – single or integrative – in the first place. To reform science we will also need to “zoom in” by formalizing our expectations within a theoretical framework and foregrounding the conditionality of our inferences.

**Acknowledgment.** The author thanks Andrew Gelman for comments on a draft.


**Financial support.** This work was supported by the National Science Foundation (CISE Nos. 2211939 and 1930642) and a Microsoft Research Faculty Fellowship.

**Competing interest.** None.

### References

- Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, 5(8), 980–989.
- Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3), 200805.
- Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 12(1), 46–61.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195–244.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3), 221–229.
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26, 1596–1618.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122.

## Discovering the unknown unknowns of research cartography with high-throughput natural description

Tanay Katiyar<sup>a</sup>, Jean-François Bonnefon<sup>b</sup>,  
Samuel A. Mehr<sup>c,d\*</sup>  and Manvir Singh<sup>e</sup>

<sup>a</sup>Institut Jean Nicod, Département d'études cognitives, École normale supérieure (ENS-PSL), Paris, France; <sup>b</sup>Toulouse School of Economics, Centre National de la Recherche Scientifique (TSM-R), Toulouse, France; <sup>c</sup>School of Psychology, University of Auckland, Auckland, New Zealand; <sup>d</sup>Yale Child Study Center, Yale University, New Haven, CT, USA and <sup>e</sup>Department of Anthropology, University of California-Davis, Davis, CA, USA

[tanay.katiyar20@gmail.com](mailto:tanay.katiyar20@gmail.com)

[jean-francois.bonnefon@tse-fr.eu](mailto:jean-francois.bonnefon@tse-fr.eu); <https://jfbonnefon.github.io>

[sam@yale.edu](mailto:sam@yale.edu); <https://mehr.nz/>

[manvir.manvir@gmail.com](mailto:manvir.manvir@gmail.com); <https://manvir.org>

\*Corresponding author.

doi:10.1017/S0140525X23002170, e50

### Abstract

To succeed, we posit that research cartography will require high-throughput natural description to identify unknown unknowns in a particular design space. High-throughput natural description, the systematic collection and annotation of representative corpora of real-world stimuli, faces logistical challenges, but these can be overcome by solutions that are deployed in the later stages of integrative experiment design.

The integrative approach advocated by Almaatouq et al. starts with mapping a research field onto an  $n$ -dimensional design space that defines the universe of relevant experiments – what they call “research cartography” (target article, sect. 3.1 para. 2). They suggest that the design space’s dimensions can be extracted from available taxonomies, prior experimental research, and practical experience. However, as they acknowledge, this approach is vulnerable to unknown unknowns: Taxonomies, prior experiments, and practical experience may all fail to identify important dimensions which should be included in the design space.

Here, we focus on one way of identifying unknown unknowns: High-throughput natural description. This approach may help research cartographers to uncover missing dimensions of the research design space, at a cost comparable to the later stages of the integrative experiment design.

To appreciate the value of high-throughput natural description, consider cases where researchers noticed a discrepancy between the experimental stimuli and the naturalistic variation of these stimuli. For instance, Schutz and Gillard (2020) showed that many experiments studying nonspeech auditory perception used flat tones as stimuli, despite the fact that such tones are unrealistic: Their content lacks dynamic changes found in the temporal structure of naturalistic sounds. Experiments that included such naturalistic content made novel discoveries about the auditory system. For example, a study of audiovisual integration showed that tones with a temporal structure similar to impact sounds, like the sound of a xylophone, but not flat tones, which lack temporal variation, were reliably integrated with visual

information when participants judged tone duration (Schutz & Kubovy, 2009).

Similarly, Dawel, Miller, Horsburgh, and Ford (2021) and Barrett, Adolphs, Marsella, Martinez, and Pollak (2019) showed that many experiments studying face perception used highly standardised and posed facial configurations which are not representative of the real-world variation in facial configurations. When naturalistic facial configurations are used in experiments, reported findings differ from previous results. For example, using naturalistic facial stimuli, Sutherland et al. (2013) found that facial first impressions have three underlying dimensions (trustworthiness, dominance, and youthfulness/attractiveness) instead of just two (trustworthiness and dominance), as previously reported when standardised facial stimuli were used (Oosterhof & Todorov, 2008; Todorov, Said, Engell, & Oosterhof, 2008).

In these examples, researchers noticed and resolved some discrepancy between the variation of experimental and real-world stimuli. Such an approach, while useful, does not completely solve the problem of unknown unknowns. This is because there may be many more real-world variations in stimuli that could update one’s understanding of a phenomenon, if they were introduced in experimental designs. However, a researcher cannot identify them unless they have a thorough description of real-world variation.

One solution to this issue is “high-throughput natural description”: *The systematic collection and annotation of large, representative corpora of real-world stimuli to identify unknown unknowns.*

An example in the field of emotion perception demonstrates the value of this approach. By collecting and annotating 7 million pictures of faces and 10,000 hours of filmed video from the internet, Srinivasan and Martinez (2018) discovered that the emotion-category labels of disgust, anger, sadness, and happiness are associated with 1, 5, 5, and 17 “distinct” facial configurations, respectively. Such variation in the range of facial configurations conveying different emotions was an unknown unknown in the research cartography of emotion perception, and studies investigating responses to facial configurations expressing certain emotion categories have yet to investigate responses to the entirety of the observed variation, to the best of our knowledge (Barrett et al., 2019). Thus, high-throughput natural description can aid in defining the design space of relevant experiments via the identification of unknown unknowns.

However, this solution is not an easy fix to the problem of unknown unknowns. Large-scale naturalistic observation is logistically challenging. Obtaining 7 million images of faces from the internet is in itself difficult, but the difficulty ramps up if researchers wish to obtain a sample of faces from more diverse sources. Furthermore, large-scale annotation can be as challenging as large-scale naturalistic observation. For example, creating a corpus of 7 million faces that is useful for answering different research questions requires annotating the images for meaningful dimensions. Coding action units (specific facial muscle movements) manually via human annotators in these images can require expertise, or can take years when the dataset is extremely large (Benitez-Quiroz, Srinivasan, & Martinez, 2016; Srinivasan & Martinez, 2018). Furthermore, the pool of annotators must itself be (very) large, not only to deal with the size of the corpus, but also to identify relevant individual and cultural variations in the way coders perceive the dimensionality of the stimuli.

In sum, while high-throughput natural description aids in the identification of unknown unknowns of a research design

space, it introduces significant logistical challenges. However, these challenges can be surmounted via a combination of *mass collaboration*, *automation* (a use case is already present in the aforementioned emotion perception example where Srinivasan & Martinez, 2018, use a computer vision algorithm to annotate action units in the internet images; Benitez-Quiroz et al., 2016; Yitzhak et al., 2017), *citizen science* (Awad et al., 2018, 2020; Hilton & Mehr, 2021), and *gamification* (Long, Simson, Buxó-Lugo, Watson, & Mehr, 2023). In fact, Almaatouq et al. already propose that these aforementioned solutions could be deployed in the later stages of the integrative experiment design.

Nonetheless, the application of these solutions for executing high-throughput natural description should not be ignored, as they amplify concerns about the up-front costs and inclusivity of the integrative approach. Few research groups may have the resources to implement an integrative experiment design, and fewer groups still may be able to solve its unknown unknowns problem during the research cartography stage. While we are enthusiastic about the ideas in the target article, we believe it is necessary to be explicit and constructive about the requirements of an integrative experiment design approach.

**Acknowledgments.** T. K. would like to thank Dr. Julie Grèzes for briefly discussing the current state of the face perception and social cognition literature.

**Financial support.** S. A. M. is supported by NIH DP5OD024566. J.-F. B. acknowledges support from grants ANR-19-PI3A-0004, ANR-17-EURE-0010, and the research foundation TSE-Partnership.


**Competing interest.** None.

## References

- Awad, E., Dsouza, S., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, 63(3), 48–55.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563, 59–64.
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68. <https://doi.org/10.1177/1529100619832930>
- Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In 2016 IEEE Conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA (pp. 5562–5570). <https://doi.org/10.1109/CVPR.2016.600>
- Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2021). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, 54(4), 1889–1901. <https://doi.org/10.3758/s13428-021-01705-3>
- Hilton, C., & Mehr, S. (2021). Citizen science can help to alleviate the generalizability crisis. 45, e21.
- Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., & Mehr, S. A. (2023). How games can make behavioural science better. *Nature*, 613(7944), 433–436.
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(32), 11087–11092. <https://doi.org/10.1073/pnas.0805664105>
- Schutz, M., & Gillard, J. (2020). On the generalization of tones: A detailed exploration of non-speech auditory perception stimuli. *Scientific Reports*, 10(1), 9520. <https://doi.org/10.1038/s41598-020-63132-2>
- Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6), 1791.
- Srinivasan, R., & Martinez, A. M. (2018). Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Transactions on Affective Computing*, 12(3), 707–721.
- Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, 127(1), 105–118. <https://doi.org/10.1016/j.cognition.2012.12.001>
- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, 12(12), 455–460. <https://doi.org/10.1016/j.tics.2008.10.001>

Yitzhak, N., Giladi, N., Gurevich, T., Messinger, D. S., Prince, E. B., Martin, K., & Aviezer, H. (2017). Gently does it: Humans outperform a software classifier in recognizing subtle, nonstereotypical facial expressions. *Emotion*, 17(8), 1187–1198. <https://doi.org/10.1037/emo0000287>

## Against naïve induction from experimental data

David Kellen<sup>a\*</sup> , Gregory E. Cox<sup>b</sup>, Chris Donkin<sup>c</sup>,  
John C. Dunn<sup>d</sup> and Richard M. Shiffrin<sup>e</sup>

<sup>a</sup>Department of Psychology, Syracuse University, Syracuse, NY, USA;

<sup>b</sup>Department of Psychology, College of Arts and Sciences, University at Albany, State University of New York, Albany, NY, USA; <sup>c</sup>Department of Psychology, Ludwig Maximilian University of Munich, München, Germany; <sup>d</sup>Department of Psychology, University of Western Australia, Perth, WA, Australia and

<sup>e</sup>Psychological and Brain Sciences Department, Indiana University, Bloomington, IN, USA

[davekellen@gmail.com](mailto:davekellen@gmail.com)

[gecox@albany.edu](mailto:gecox@albany.edu)

[christopher.donkin@gmail.com](mailto:christopher.donkin@gmail.com)

[john.dunn@uwa.edu.au](mailto:john.dunn@uwa.edu.au)

[shiffrin@indiana.edu](mailto:shiffrin@indiana.edu)

\*Corresponding author.

doi:10.1017/S0140525X2300211X, e51

### Abstract

This commentary argues against the indictment of current experimental practices such as piecemeal testing, and the proposed integrated experiment design (IED) approach, which we see as yet another attempt at automating scientific thinking. We identify a number of undesirable features of IED that lead us to believe that its broad application will hinder scientific progress.

After so many years observing the prosecution of *p*-values and everyday laboratory life, we are pleased to see a growing number of researchers turning their attention to critical matters such as theory development and experimentation (e.g., Proulx & Morey, 2021). But as we transition into these important new debates, it is crucial to avoid past intellectual excesses. In particular, we note a tendency to embrace passive technological solutions to problems of scientific inference and discovery that make little room for the kind of active theory building and critical thinking that in fact result in meaningful scientific advances (see Singmann et al., 2023). In this vein, we wish to express serious reservations regarding Almaatouq et al.'s critique.

The observation of puzzling, incongruent, and incommensurate results across studies is a common affair in the experimental sciences (see Chang, 2004; Galison, 1987; Hacking, 1983). Indeed, one of the central roles of experimentation is to “create, produce, refine and stabilize phenomena” (Hacking, 1983, p. 229), which is achieved through an iterative process that includes the ongoing improvement of experimental apparatus (see Chang, 2004; Trendler, 2009) and relevant variables (Jantzen, 2021). This process was discussed long ago by Maxwell (1890/1965), who described it as removing the influence of “disturbing agents” from a “field of investigation.”



Looking back at the history of modern memory research, we can identify this process in the development of experimental tasks (e.g., recognition, cued recall) with clear procedures (study/test phases) and stimuli (e.g., high-frequency words). This process is also manifest in the resolution of empirical puzzles, such as the innumerable exceptions, incongruencies, and boundary conditions encountered by researchers in the search for the “laws of memory” (for a review, see Roediger, 2008). Far from insurmountable, these empirical puzzles have been continuously resolved through the interplay of tailored experiments and theories (e.g., Cox & Shiffrin, 2017; Hotaling, Donkin, Jarvstad & Newell, 2022; Humphreys, Bain, & Pike, 1989; Roediger & Blaxton, 1987; Seamon et al., 1995; Turner, 2019; Vergauwe & Cowan, 2015). More specifically, candidate theories are constructed to explain existing results by postulating constructs (e.g., “trace strength”) and specifying how those constructs are related to observables (e.g., “more study time leads to more trace strength which leads to faster response times”). These theories also specify what should not be relevant, thereby identifying potential confounding variables that future experiments should control. For an exemplary case, consider the domain of short-term memory, where we can find a large body of empirical phenomena (e.g., Oberauer et al., 2018) alongside explanatory accounts that can accommodate them (e.g., interference-based theories; see Lewandowsky, Oberauer, & Brown, 2009).

Against this backdrop, it is difficult to find Almaatouq et al.’s critique convincing. On the one hand, they fail to explain the success of existing experimental practices (e.g., piecemeal testing) in domains such as human memory. On the other, their treatment case studies such as “group synergy,” which has amassed a wealth of conflicting findings, do not include any indication that the process described above has failed. This omission opens a number of possible explanations. For example, incongruent results may reflect experimental artifacts or hidden *ceteris paribus* clauses and other preconditions (Meehl, 1990, p. 109) – can we really say that these procedures have been thoroughly pursued? Alternatively, incongruent results could be a sign that those results should not be treated as part of the same “space” in the first place, that is, that they do not define a cohesive body of results that can be explained by a common theory.

Moving on to the actual proposal of integrated experiment design (IED), we find its potential contribution to be largely negative. Referring back to Maxwell’s (1890/1964) description, what IED proposes is to allow “disturbing agents” back into the “field of investigation” as long as they are appropriately tagged and recorded. It is difficult to imagine how Newton’s laws of motion could ever emerge from large-scale experiments evaluating different shapes of objects, velocities, viscosities, surface textures, and so on. Our main concerns with IED are summarized below:

- (1) By placing a premium on commensurability, IED decreases the chances of new and unexpected findings (Shiffrin, Börner, & Stigler, 2018).
- (2) By shifting researchers’ resources toward the joint observation of a large number of factors, IED disrupts the piecemeal efforts in experimentation and theorization that illuminate the processes underlying human data generation. For instance, it makes it difficult to tell an important result from one caused by a confound (for discussions, see

Garcia-Marques & Ferreira, 2011; Kellen, 2019; Shiffrin & Nobel, 1997).

- (3) IED turns existential-abductive reasoning on its head: Instead of developing explanatory constructs (e.g., model development) in response to existing covariational information, a construct would be assumed a priori in the form of an empty vessel, to be later infused by the results of an experiment manipulating factors presumably related to it. For instance, the construct “attention” would be identified with the experimental manipulations thought to be relevant to “attention.” This concern is materialized by the treatment of the so-called Moral Machine, a statistical model summarizing the observed relationships between moral judgments and a host of variables, as a bona fide theory of moral reasoning.
- (4) By introducing a large number of factors, IED can easily degrade researchers’ ability to identify which theoretical components are doing the leg work and which ones are failing, especially when compared to piecemeal testing (e.g., Birnbaum, 2008; Dunn & Rao, 2019; Kellen, Steiner, Davis-Stober, & Pappas, 2020). The recent application of IED to risky-choice modeling (Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021) illustrates this concern, as it is unclear which specific circumstances are leading one choice model to outperform another (e.g., is context dependency driven by feedback?).

It is our judgment that there is no one best way to do science, and that attempts to tell scientists how to do their job, including IED, will slow and hinder progress. IED is solving a problem that does not exist and introduces a problem that science should do without.

**Financial support.** David Kellen was supported by NSF CAREER Award ID 2145308.

**Competing interest.** None.

## References

- Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, 115, 463–501.
- Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
- Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review*, 124, 795–860.
- Dunn, J. C., & Rao, L. L. (2019). Models of risky choice: A state-trace and signed difference analysis. *Journal of Mathematical Psychology*, 90, 61–75.
- Galison, P. L. (1987). *How experiments end*. University of Chicago Press.
- Garcia-Marques, L., & Ferreira, M. B. (2011). Friends and foes of theory construction in psychological science: Vague dichotomies, unified theories of cognition, and the new experimentalism. *Perspectives on Psychological Science*, 6, 192–201.
- Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.
- Hotaling, J. M., Donkin, C., Jarvstad, A., & Newell, B. R. (2022). MEM-EX: An exemplar memory model of decisions from experience. *Cognitive Psychology*, 138, 101517.
- Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, 96, 208–233.
- Jantzen, B. C. (2021). Scientific variables. *Philosophies*, 6, 103.
- Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, 2, 160–165.
- Kellen, D., Steiner, M. D., Davis-Stober, C. P., & Pappas, N. R. (2020). Modeling choice paradoxes under risk: From prospect theories to sampling-based accounts. *Cognitive Psychology*, 118, 101258.
- Lewandowsky, S., Oberauer, K., & Brown, G. D. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences*, 13, 120–126.

- Maxwell, J. C. (1860/1965). General considerations concerning scientific apparatus. In W. D. Niven (Ed.), *The scientific papers of James Clerk Maxwell* (Vol. 2, pp. 505–522). Dover.
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1, 108–141.
- Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., ... Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, 144, 885–958.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, 372, 1209–1214.
- Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, 16, 671–681.
- Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, 59, 225–254.
- Roediger, H. L. III, & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. In D. S. Gorfein & R. R. Hoffman (Eds.), *Memory and learning: The Ebbinghaus Centennial conference* (pp. 349–379). Erlbaum.
- Seamon, J. G., Williams, P. C., Crowley, M. J., Kim, I. J., Langer, S. A., Orne, P. J., & Wishengrad, D. L. (1995). The mere exposure effect is based on implicit memory: Effects of stimulus type, encoding conditions, and number of exposures on recognition and affect judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 711–721.
- Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 2632–2639.
- Shiffrin, R. M., & Nobel, P. A. (1997). The art of model development and testing. *Behavior Research Methods, Instruments, & Computers*, 29, 6–14.
- Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., ... Shiffrin, R. M. (2023). Statistics in the service of science: Don't let the tail wag the dog. *Computational Brain & Behavior*, 6, 64–83.
- Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, 19, 579–599.
- Turner, B. M. (2019). Toward a common representational framework for adaptation. *Psychological Review*, 126, 660–692.
- Vergauwe, E., & Cowan, N. (2015). Working memory units are all in your head: Factors that influence whether features or objects are the favored units. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1404–1416.

## Beyond integrative experiment design: Systematic experimentation guided by causal discovery AI

Erich Kummerfeld<sup>a\*</sup> and Bryan Andrews<sup>b</sup>

<sup>a</sup>Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA and <sup>b</sup>Department of Psychiatry and Behavioral Sciences, University of Minnesota, Minneapolis, MN, USA  
[erichk@umn.edu](mailto:erichk@umn.edu); <https://erichkummerfeld.com/>  
[andr1017@umn.edu](mailto:andr1017@umn.edu)

\*Corresponding author.

doi:10.1017/S0140525X23002273, e52

### Abstract

Integrative experiment design is a needed improvement over ad hoc experiments, but the specific proposed method has limitations. We urge a further break with tradition through the use of an enormous untapped resource: Decades of causal discovery artificial intelligence (AI) literature on optimizing the design of systematic experimentation.

Almaatouq et al. propose a break from tradition to accelerate scientific progress, and we applaud them for it. However, we urge an

even further shift to incorporate theory and methods from causal discovery, a subfield of machine learning with decades of research on artificial intelligence (AI)-guided causal learning and experiment design. Causal discovery has not been well leveraged in the experimental sciences perhaps because it also breaks from tradition – statistical tradition.

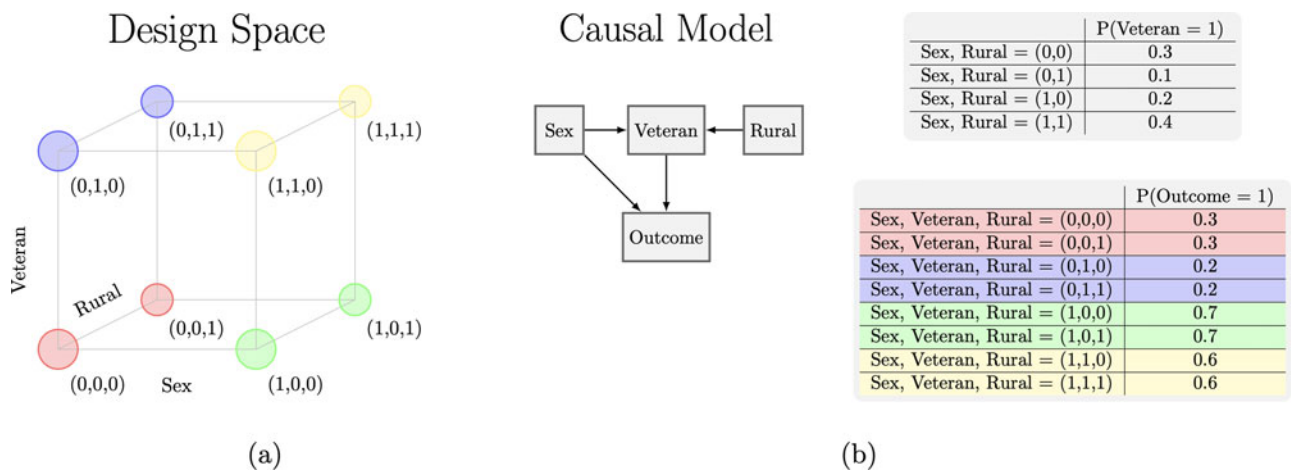
Causal discovery contains a growing collection of methods for learning multivariate structural causal models (Pearl, 2000; Spirtes et al., 2000). Design spaces can be represented as a substructure of a larger structural causal model (illustrated in Fig. 1), making causal discovery closely aligned with research cartography. It is not surprising then that some of the challenges faced by integrative experiment design might be overcome with causal discovery. We focus on three such challenges: Practical application and scalability, confined inferential scope, and unknown causal factors.

Regarding the *practical application* of design spaces, causal discovery can learn entire causal models from nonexperimental data alone, but the direction of causal relationships can be difficult to identify (Hoyer, Janzing, Mooij, Peters, & Schölkopf, 2008; Peters, Janzing, & Schölkopf, 2011; Peters et al., 2014; Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006; Shimizu et al., 2011; Spirtes et al., 2000). Causal discovery can be applied to experimental data to resolve this limitation. Multiple methods are capable of combining datasets with: Both experimental and observational samples, samples with nonidentical variables, and samples from different contexts and populations (Bareinboim & Pearl, 2016; Huang et al., 2020; Mooij, Magliacane, & Claassen, 2020; Peters, Bühlmann, & Meinshausen, 2016). Incorporating these methods could enable increased flexibility when dealing with practical study design challenges.

*Scalability* is another practical issue: The size of these spaces makes complete search infeasible. Causal discovery methods can scale to large numbers of variables, however. Even a million variables is possible (Ramsey, Glymour, Sanchez-Romero, & Glymour, 2017), but this applies to sparse models. In sparse models, each variable is directly related to only a small number of other variables. When variables have large numbers of interacting causes, causal discovery also suffers scalability problems (Spirtes et al., 2000). However, such situations may not be common in reality. Like how linear and Gaussian modeling are surprisingly effective, sparse models often capture the important elements of a causal system. As alternatives, the active learning methods Almaatouq et al. point to could be used, and active learning causal discovery methods also exist (Ghassami, Salehkaleybar, Kiyavash, & Bareinboim, 2018; Hyttinen, Eberhardt, & Hoyer, 2013a; Lindgren, Kocaoglu, Dimakis, & Vishwanath, 2018).

*Confined inferential scope* limits the kinds of information that can be learned. For example, let X, Y, and Z be variables. Some study designs allow researchers to learn that X causes Z and Y causes Z, but prevent researchers from learning whether X mediates the effect of Y on Z. In a pair of papers, Mayo-Wilson (2011, 2014) proved: (1) certain causal facts cannot be learned from a system of experiments that each only investigate a single exposure–outcome pair, (2) the proportion of unlearnable facts approaches 100% as the complexity of the system increases, and (3) overcoming this requires that each experiment measures more variables than an exposure–outcome pair. By focusing on a single experiment under different conditions, Almaatouq et al. are at risk of being confined to a space of causal facts not much greater than the ad hoc experimentation they are trying to break away from.

Researchers ought to simultaneously measure as many relevant variables as possible. This happens naturally when planning to



**Figure 1** (Kummerfeld and Andrews). (a) Hypothetical design space with three binary dimensions: Veteran status, rural status, and sex. Different experiment outcomes are colored red, green, blue, and yellow. Note that in this hypothetical example, rural status makes no difference to the outcome of the experiment, while each of the four combinations of veteran status and sex produce different outcomes. (b) A causal model that would correspond to the example design space. The structure of the causal model is shown on the left, and the two causal dependency tables are shown on the right: One for veteran status, which depends on sex and rural, and the other for outcome. The table for outcome is shown with rural included, to make the comparison with the design space clear, but in a normal causal model rural would not be included in this table as no arrow points directly from rural to outcome in the model structure.

use causal discovery methods. Most causal discovery methods treat all variables equally, with no labeled outcome variable. It is normal in causal discovery to cast a wide net and use measurements from a larger number of variables, and then simultaneously model them with an algorithm. There is a growing body of papers applying this approach, including some in the social and behavioral sciences (Bronstein, Everaert, Kummerfeld, Haynos, & Vinogradov, 2022a; Bronstein, Kummerfeld, MacDonald, & Vinogradov, 2022b; Shen, Ma, Vemuri, & Simon, 2020; Stevenson et al., 2022).

Unknown causal factors are ubiquitous in science and, unbeknownst to the researcher, can modify the context under which the data were collected. This commonly manifests as latent confounding. In the integrative experimental design paradigm it would occur as a failure to fully specify the design space. Research cartography could possibly solve this, but it is unclear how.

In contrast, causal discovery offers multiple solutions to unknown causal factors. Many causal discovery algorithms are only correct assuming “causal sufficiency”: That there are no unknown causal factors causing two or more measured variables. However there are also many papers developing theory and methods without assuming causal sufficiency (Chen et al., 2021; Hyttinen, Hoyer, Eberhardt, & Jarvisalo, 2013b; Ogarrio, Spirtes, & Ramsey, 2016; Spirtes et al., 2000; Zhang, 2008). In many cases the presence or absence of unknown causal factors can be identified from measured data, and there are even causal discovery methods designed to learn the causal relationships among them (Huang, Low, Xie, Glymour, & Zhang, 2022; Kummerfeld & Ramsey, 2016; Xie et al., 2022).

Unfortunately, causal discovery has had limited application in the experimental sciences. We hope this commentary helps to raise awareness of these resources. Almaatouq et al. make it clear that there is a demand for these research products in the social and behavioral sciences. There is a serious barrier to the adoption and use of causal discovery: Much of it is buried and scattered among journals covering relatively unapplied topics such as theoretical machine learning and philosophy of science.

We expect that in the future causal discovery will gain presence in journals on experimental methods and design or topics such as behavioral and brain sciences.

**Financial support.** E. K. was supported by funding through Grant No. NCR11TR002494-01 and B. A. was supported by T32 DA037183. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interest.** None.

### References

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27), 7345–7352.

Bronstein, M. V., Everaert, J., Kummerfeld, E., Haynos, A. F., & Vinogradov, S. (2022a). Biased and inflexible interpretations of ambiguous social situations: Associations with eating disorder symptoms and socioemotional functioning. *International Journal of Eating Disorders*, 55(4), 518–529. <https://doi.org/10.1002/eat.23688>

Bronstein, M. V., Kummerfeld, E., MacDonald, A., III, & Vinogradov, S. (2022b). Willingness to vaccinate against SARS-CoV-2: The role of reasoning biases and conspiracist ideation. *Vaccine*, 40(2), 213–222.

Chen, W., Zhang, K., Cai, R., Huang, B., Ramsey, J., Hao, Z., & Glymour, C. (2021). FRITL: A hybrid method for causal discovery in the presence of latent confounders. *arXiv [cs.LG]*. <http://arxiv.org/abs/2103.14238>

Ghassami, A., Salehkaleybar, S., Kiyavash, N., & Bareinboim, E. (2018). Budgeted experiment design for causal structure learning. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 1724–1733). PMLR.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21, 689–696. <https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcd94a4b6-Abstract.html>

Huang, B., Low, C. J. H., Xie, F., Glymour, C., & Zhang, K. (2022). Latent hierarchical causal structure discovery with rank constraints. *Advances in Neural Information Processing Systems*, 35, 5549–5561.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., & Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research: JMLR*, 21(1), 3482–3534.

Hyttinen, A., Eberhardt, F., & Hoyer, P. O. (2013a). Experiment selection for causal discovery. *Journal of Machine Learning Research: JMLR*, 14, 3041–3071.

Hyttinen, A., Hoyer, P. O., Eberhardt, F., & Jarvisalo, M. (2013b). Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* (pp. 301–310).



- Kummerfeld, E., & Ramsey, J. (2016). Causal Clustering for 1-Factor Measurement Models. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1655–1664). <https://doi.org/10.1145/2939672.2939838>
- Lindgren, E., Kocaoglu, M., Dimakis, A. G., & Vishwanath, S. (2018). Experimental design for cost-aware learning of causal graphs. *Advances in Neural Information Processing Systems*, 31, 5284–5294. <https://proceedings.neurips.cc/paper/2018/hash/ba3e9b6a519cfddc560b5d53210df1bd-Abstract.html>
- Mayo-Wilson, C. (2011). The problem of piecemeal induction. *Philosophy of Science*, 78(5), 864–874.
- Mayo-Wilson, C. (2014). The limits of piecemeal causal inference. *The British Journal for the Philosophy of Science*, 65(2), 213–249.
- Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research: JMLR*, 21(1), 3919–4026.
- Ogarrio, J. M., Spirtes, P., & Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. *JMLR Workshop and Conference Proceedings*, 52, 368–379.
- Pearl, J. (2000). *Causality: Models, reasoning and inference* (Vol. 29). Springer.
- Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 78(5), 947–1012.
- Peters, J., Janzing, D., & Schölkopf, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12), 2436–2450.
- Peters, J., Mooij, J., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15, 2009–2053. <https://www.jmlr.org/papers/volume15/peters14a/peters14a.pdf>
- Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2), 121–129.
- Shen, X., Ma, S., Vemuri, P., & Simon, G., & Alzheimer's Disease Neuroimaging Initiative. (2020). Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. *Scientific Reports*, 10(1), 2975.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research: JMLR*, 7 (Oct), 2003–2030.
- Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., ... Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *The Journal of Machine Learning Research*, 12, 1225–1248. <https://www.jmlr.org/papers/volume12/shimizul1a/shimizul1a.pdf>
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT Press.
- Stevenson, B. L., Kummerfeld, E., Merrill, J. E., Blevins, C., Abrantes, A. M., Kushner, M. G., & Lim, K. O. (2022). Quantifying heterogeneity in mood-alcohol relationships with idiographic causal models. *Alcoholism, Clinical and Experimental Research*, 46(10), 1913–1924.
- Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., & Zhang, K. (2022). Identification of linear non-Gaussian latent hierarchical structure. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 24370–24387). PMLR.
- Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16), 1873–1896.

## Abstract

The target article argues researchers should be more ambitious, designing studies that systematically and comprehensively explore the space of possible experiments in one fell swoop. We argue that while “systematic” is rarely achievable, “comprehensive” is often enough. Critically, the recent popularization of massive online experiments shows that comprehensive studies are achievable for most cognitive and behavioral research questions.

Almaatouq et al. provide an incisive and welcome critique of the dominant one-at-a-time paradigm. They argue for integrative studies that systematically and comprehensively explore the “universe of possible experiments” (target article, sect. 2.2, para. 1) in each domain of inquiry. While we are sympathetic to the goal, Almaatouq et al. overemphasize *systematic* at the expense of *comprehensive*.

As we see it, the core problem with the one-at-a-time approach is that it is too slow. It is not news that most studies extrapolate broadly from a miniscule sample of stimuli, subject demographics, and experimental paradigms, resulting in a long-running generalizability crisis (Clark, 1973; Henrich, Heine, & Norenzayan, 2010; Judd, Westfall, & Kenny, 2012; Yarkoni, 2022). Even large literatures often fail to do much more than scratch the surface of possibilities (e.g., Hartshorne & Snedeker, 2013; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021). It is as if we set out to explore the universe of possible experiments, but spent most of our time hanging out at the hotel pool.

In principle, Almaatouq et al.'s integrative experiment approach is ideal: Find the set of parameters that describe the universe of possible experiments and then survey systematically. Unfortunately, this requires better understanding of the phenomenon than we usually have. Indeed, the cognitive and behavioral sciences remain largely in Kuhn's preparadigmatic phase (Kuhn, 2012), characterized by conflicting and incommensurate theories, each with its own set of assumptions, methods, and observations.

Let us illustrate the difficulty with an easy-to-articulate question: Are children better at learning all aspects of syntax or just certain parts? Answering this question requires comparing how quickly older and younger learners learn each component of syntax. The problem is that different theories propose radically different visions of what syntax is and how it might be subdivided. Theories differ in terms of whether syntax is governed by large numbers of highly articulated, abstract rules that combine structurally simple words; by a small number of simple rules that combine internally complex words; or by superimposed, prototype-like patterns, with no distinction between words and rules; among other possibilities (Chomsky, 2014; Goldberg, 1995, 2009; Hopcroft, Motwani, & Ullman, 2001; Steedman, 2001). Even where theorists agree on the structure, they disagree on processing, with the same grammatical patterns subserved by different cognitive/neurological systems at different times or by different people (O'Donnell, 2015; Ullman, 2001).

To make matters worse, none of these are complete theories that can be applied to arbitrary stimuli. For starters, predictions depend on ancillary assumptions that are left as open empirical questions (such as the relative preference for generalizations vs.

## Don't let perfect be the enemy of better: In defense of unparameterized megastudies

Wei Li and Joshua K. Hartshorne\* 

Department of Psychology & Neuroscience, Boston College, Chestnut Hill, MA, USA

[liacz@bc.edu](mailto:liacz@bc.edu), <https://www.weiir.xyz/>

[joshua.hartshorne@hey.com](mailto:joshua.hartshorne@hey.com), <https://l3atbc.org>

\*Corresponding author.

doi:10.1017/S0140525X23002121, e53

one-off computations). More problematically, different theories often prioritize explaining distinct phenomena (common or rare utterances; highly productive patterns or semi-idiomatic expressions; early child language or mature usage; similarities across languages or cross-linguistic differences); one theory may not make clear predictions about the core motivating phenomena for another, and vice versa. While we believe this reflects the difficulty of the problem, not the diligence of the researchers, the outcome is the same: There is a lot of theoretical progress lying between here and the integrative experiments proposed by Almaatouq et al.

Even our example above underestimates the problem. Almaatouq et al. focus primarily on sampling from a stimulus space, not a task space. Two of their examples (trolley problems and risky-choice scenarios) are narrowly defined paradigms for studying much broader phenomena (moral reasoning and decision making under uncertainty). Their third example (masked cueing) does involve manipulating some task parameters beyond the stimuli themselves, but remains tied to a narrowly circumscribed task.

This might be fine if we fully understood the relationship between tasks and the underlying cognitive processes, but mostly we do not. Consider, for instance, measures of cognitive control – itself one of the most thoroughly investigated constructs in cognitive psychology. There are a number of popular tasks used to study cognitive control, including the masked cuing paradigm described by Almaatouq et al. Recently, one of us directly compared cognitive control as measured by three closely related tasks: The Simon, Stroop, and flanker tasks (Erb et al., 2023). Two massive online experiments with more than 20,000 participants revealed that these three tasks show strikingly different patterns of change in performance over the lifespan and near-zero correlations. Thus, integrative studies of cognitive control need to sample not just across stimuli but also paradigms. However, it is not clear that the differences across paradigms/tasks can be easily parameterized. Indeed, advances in our fields often owe themselves to the creation of new paradigms that open up new questions or comparisons.

Perhaps we are too pessimistic. Perhaps most questions resemble trolley problems and few resemble syntax or cognitive control (though we note that one of Almaatouq et al.'s three examples actually investigated cognitive control). But we would hate to predicate moving beyond the one-at-a-time approach on the widespread feasibility of parameterization. We worry that this licenses researchers (and editors and funders) to let perfect be the enemy of better – because we can do much better.

In particular, Almaatouq et al. may have only been able to find three examples of *systematic* exploration of the universe of possible experiments, but *comprehensive* explorations abound. This includes megastudies that test large, diverse sets of stimuli (e.g., Breithaupt, Li, & Kruschke, 2022; Brysbaert, Stevens, Mandra, & Keuleers, 2016; De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019; Hartshorne, Bonial, & Palmer, 2014), broad subject demographics (e.g., Bleidorn et al., 2013; Hartshorne, Tenenbaum, & Pinker, 2018; Nosek, Banaji, & Greenwald, 2002; Riley et al., 2016; Soto, John, Gosling, & Potter, 2011; Spiers, Coutrot, & Hornberger, 2023), or a range of related tasks (e.g., Erb, Germine, & Hartshorne, 2023; Hampshire, Highfield, Parkin, & Owen, 2012). Even without systematic exploration, these studies have produced major theoretical discoveries. They

have also been instrumental in identifying important Almaatouq et al.-style parameters for subsequent systematic exploration. Critically, as Almaatouq et al. explain, the technology exists to conduct megastudies for most cognitive and behavioral questions, typically at lower aggregate cost than the status quo (see also Gosling & Mason, 2015; Li, Germine, Mehr, Srinivasan, & Hartshorne, 2022; Long, Simson, Buxó-Lugo, Watson, & Mehr, 2023). In short, our critique is of the “yes, and” variety. Yes, conduct systematic integrative metastudies when you can. And, when you cannot, conduct less systematic megastudies.

**Financial support.** The authors acknowledge funding from NSF 2229631 to J. K. H.


**Competing interest.** None.

## References

- Bleidorn, W., Klimstra, T. A., Denissen, J. J., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world: A cross-cultural examination of social-investment theory. *Psychological Science*, 24(12), 2530–2540.
- Breithaupt, F., Li, B., & Kruschke, J. K. (2022). Serial reproduction of narratives preserves emotional appraisals. *Cognition and Emotion*, 36(4), 581–601.
- Brysbaert, M., Stevens, M., Mandra, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7, 1116.
- Chomsky, N. (2014). *The minimalist program*. MIT Press.
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4), 335–359.
- De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The “small world of words” English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51, 987–1006.
- Erb, C. D., Germine, L., & Hartshorne, J. K. (2023). Fractionating cognitive control: Congruency tasks reveal divergent developmental trajectories. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001429>
- Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
- Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, 20(1), 93–127.
- Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, 66, 877–902.
- Hampshire, A., Highfield, R. R., Parkin, B. L., & Owen, A. M. (2012). Fractionating human intelligence. *Neuron*, 76(6), 1225–1237.
- Hartshorne, J. K., Bonial, C., & Palmer, M. (2014). The VerbCorner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. In Proceedings of the 52nd annual meeting of the association for computational linguistics (vol. 2: Short Papers, pp. 397–402). Baltimore, Maryland, USA, 23–25 June 2014.
- Hartshorne, J. K., & Snedeker, J. (2013). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes*, 28(10), 1474–1508.
- Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177, 263–277.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83.
- Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2001). Introduction to automata theory, languages, and computation. *ACM Sigact News*, 32(1), 60–65.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54.
- Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago Press.
- Li, W., Germine, L. T., Mehr, S. A., Srinivasan, M., & Hartshorne, J. K. (2022). Developmental psychologists should adopt citizen science to improve generalization and reproducibility. *Infant and Child Development*, e2348.
- Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., & Mehr, S. A. (2023). How games can make behavioural science better. *Nature*, 613(7944), 433–436.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1), 101.

- O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science (New York, N.Y.)*, 372(6547), 1209–1214.
- Riley, E., Okabe, H., Germine, L., Wilmer, J., Esterman, M., & DeGutis, J. (2016). Gender differences in sustained attentional control relate to gender inequality across countries. *PLoS ONE*, 11(11), e0165100.
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, 100(2), 330.
- Spiers, H. J., Coutrot, A., & Hornberger, M. (2023). Explaining world-wide variation in navigation ability from millions of people: Citizen science project Sea Hero Quest. *Topics in Cognitive Science*, 15(1), 120–138.
- Steedman, M. (2001). *The syntactic process*. MIT Press.
- Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, 30, 37–69.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1.

## Is generalization decay a fundamental law of psychology?

David R. Mandel\* 

Defence Research and Development Canada and York University, Toronto, ON, Canada

[drmandel66@gmail.com](mailto:drmandel66@gmail.com)

<https://sites.google.com/site/themandelian/home>

\*Corresponding author.

doi:10.1017/S0140525X23002352, e54

### Abstract

Generalizations strengthen in traditional sciences, but in psychology (and social and behavioral sciences, more generally) they decay. This is usually viewed as a problem requiring solution. It could be viewed instead as a law-like phenomenon. Generalization decay cannot be squelched because human behavior is metastable and all behavioral data collected thus far have resulted from a thin sliver of human time.

*Generalizations decay.*

Lee J. Cronbach (1975, p. 122)

In traditional scientific disciplines, to use Scriven's (1956) terminology, generalizations generally strengthen and can even trigger productive theoretical upheavals. The partial overthrow of Newton's mechanics by Einstein's special theory of relativity would not have happened had Einstein not generalized the Galilean principle of relativity in mechanics (also apparent in Newton's theory) to electrodynamics. Generalization strengthens understanding, sometimes at great cost to status-quo theories. In contrast, in psychology – and more broadly, the social and behavioral sciences – generalizations decay.

The “problem” of generalization decay has long been the subject of scholarly attention, yet no satisfactory solution has been found (Cronbach, 1975; Scriven, 1956). A widely discussed reason for such decay is that behavioral phenomena are interactively determined but psychological theories invariably underspecify

the full range of interactions and often misspecify the nature of effects (Campbell, 1957; Yarkoni, 2022). An optimistic view is that psychology can overcome generalization decay by adopting an interactionist approach to theory generation and testing (e.g., Cronbach, 1957; Eysenck, 1997). Calls for methodological reform such as Almaatouq et al.'s proposed “integrative experiment design” also fall into the optimist's camp. However, none of the optimists' proposals confront the fact that for most topics, psychology does not offer a theoretical basis for knowing how high the order of interactions must be for generalization decay to be squelched.

Take Almaatouq et al.'s example where the design space (i.e., the space defined by all of the measured independent variables and putative moderators) has upward of 50 factors. Even if each factor was binary, the design space would have over a quadrillion cells. Almaatouq et al. cryptically refer to statistical methods that could start with a highly circumscribed sample of cells in this overwhelming space, but these statements do little to inspire confidence in the overall project.

Permitting less ambition, assume a design space of a mere dozen variables split equally between binary and ternary factors. This space has 46,656 cells. Imagine that the researchers learn that, of the 5,667 possible interaction effects, several hundred are significant including over a dozen with  $n > 7$ . Would anyone have reason for confidence that such higher-order interactions would replicate if such a costly experiment could ever be repeated? Even if they were all replicable (an amazingly improbable occurrence), would they productively advance fundamental theory in psychology? After all, Newell's (1973) concern (the entry point for Almaatouq et al.) was not mainly about the lack of generalizability in psychology but, rather, about the slim prospect of theoretical unification even among fine examples of work by the best and brightest minds of his time.

Generalization decay cannot be eliminated through design mandates or interactionist projects because human behavior is metastable over time (Gergen, 1973). Psychologists often make claims about human behavior and cognition as if it applied to all humans across time, but this is unknowable since virtually all research participant data have been collected in a sliver of human history, and even historical records only go back thousands, not hundreds of thousands, of years, as would be required. However, even if psychologists fully exploited historical records of “dead minds,” as Atari and Henrich (2023) call for, or even if they miraculously recovered the full record of humanity's past, we have no trace of humanity's actual future. We do not know what proportion of human existence lies ahead or what metastable states it will occupy. Our theories do not project us clearly into the future as in physics, which provides a basis for estimation of physical transformations of the universe over unimaginable timescales (Dyson, 1979; Krauss & Starkman, 2000). They do not even project us as well into the past. We have no equivalent of the cosmic microwave background.

Psychologists cannot study the moderating role of social factors that do not yet exist. When life expectancy is universally less than 100 years, psychologists cannot formulate a generalizable theory of lifespan development that applies to humans who might live in an epoch following actuarial escape velocity – the point at which *remaining* life expectancy *increases* with time due to mortality rates that plummet due to disruptive scientific and technological breakthroughs (de Grey, 2004). We do not know what human experience will be like in the future. Imagine that in 50



years, the interpretation of quantum mechanics (based on evidence or insights that currently do not exist) indisputably favors the Everettian many worlds hypothesis and it becomes common knowledge that each of us exists in possibly infinite branches of decohered worlds – *duh!* – what then will social psychologists have to say about the self concept?

Psychology's uncertainty about humanity's past and future may be inevitable, but its comfort with a focus on the moving present reveals a parochial disposition that the traditional sciences outgrew long ago. If young Einstein did not stretch his mind to imagine whether light waves would appear to him to be at rest if he were able to run alongside at the speed of light, and if he did not have Maxwell's equations to show that the wave-at-rest counterfactual could not resolve the equations, he may not have discovered one of the most important theories in the history of science. Psychology banished introspection as a reliable method long ago; and it does not have the equivalent of Maxwell's equations, but where are its creative Machian *Gedankenexperiments* that may lead us out of musty local minima? The absence of thought experiments that revolutionize theoretical understanding in psychology is itself a mystery that deserves scholarly attention.

Returning to Scriven, "science has not advanced by solving all problems but often by abandoning them..." (1956, p. 339). If psychology pines for generalizability, resisting the apparent law of generalization decay, psychologists will need to seek new problems and ways of understanding. A productive path forward may be to seek greater consilience with traditional sciences (Wilson, 1998). Biocosmology offers a promising recent example (Cortés, Kauffman, Liddle, & Smolin, 2022). Alternatively, psychology could accept historicism as a metatheoretical foundation (Gergen, 1973), and its future might split along such lines.

**Financial support.** This work was funded by Canadian Safety and Security Program Project CSSP-2018-TI-2394.

**Competing interest.** None.



**Note.** His Majesty the King in Right of Canada as represented by Department of National Defence.

## References

- Atari, M., & Henrich, J. (2023). Historical psychology. *Current Directions in Psychological Science*, 32(2), 176–183. <https://doi.org/10.1177/09637214221149737>
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. <https://doi.org/10.1037/h0040950>
- Cortés, M., Kauffman, S. A., Liddle, A. R., & Smolin, L. (2022). Biocosmology: Biology from a cosmological perspective. <https://doi.org/10.48550/arXiv.2204.09379>
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12(11), 671–684.
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30(2), 116–127.
- de Grey, A. D. N. J. (2004). Escape velocity: Why the prospect of extreme human life extension matters now. *PLoS Biology*, 2(6), e187.
- Dyson, F. J. (1979). Time without end: Physics and biology in an open universe. *Reviews of Modern Physics*, 51(3), 447–460.
- Eysenck, H. J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and Social Psychology*, 73(6), 1224–1237.
- Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, 26(2), 309–320.
- Krauss, L. M., & Starkman, G. D. (2000). Life, the universe, and nothing: Life and death in an ever-expanding universe. *The Astrophysical Journal*, 531, 22–30.

- Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing: Proceedings of the eighth annual Carnegie symposium on cognition* (pp. 283–310). Academic Press.
- Scriven, M. (1956). A possible distinction between traditional scientific disciplines and the study of human behavior. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science* (Vol. 1, pp. 330–339). University of Minnesota Press.
- Wilson, E. O. (1998). *Consilience: The unity of knowledge*. Vintage.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45(e1), 1–78. doi:10.1017/S0140525X20001685

## Sampling complex social and behavioral phenomena

Henrik Olsson<sup>a,b</sup>  and Mirta Galesic<sup>a,b\*</sup> 

<sup>a</sup>Complexity Science Hub, Vienna, Austria and <sup>b</sup>Santa Fe Institute, Santa Fe, NM, USA

[olsson@santafe.edu](mailto:olsson@santafe.edu); [galesic@santafe.edu](mailto:galesic@santafe.edu)

<https://www.santafe.edu/people/profile/henrik-olsson>; <https://www.santafe.edu/people/profile/mirta-galesic>

\*Corresponding author.

doi:10.1017/S0140525X23002327, e55

### Abstract

We comment on the limits of relying on prior literature when constructing the design space for an integrative experiment; the adaptive nature of social and behavioral phenomena and the implications for the use of theory and modeling when constructing the design space; and on the challenges of measuring random errors and lab-related biases in measurement without replication.

We welcome this thoughtful and creative set of ideas for improving experimentation in the social sciences. We offer several points for discussion that might further clarify and strengthen the authors' arguments.

First, how should the design space be constructed? The authors suggest that the design space from which researchers can sample various aspects of the phenomena of interest can be constructed mostly by reviewing past literature. However, past studies are often a biased sample of the phenomena of interest, driven by implicit or explicit theories their authors had at the time, by methodological limitations, or an adherence to a particular experimental paradigm.

An example from the judgment and decision-making literature is the phenomenon of overconfidence. The assumption that an experimenter can choose "good general knowledge items" led to results suggesting that people almost always show overconfidence. But using the Brunswikian ideas of representative design, later studies (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994) showed that the items that had been previously selected were not representative of the whole population of items people experience in the real world. By randomly sampling from the whole population of items, which approximates representative design, studies showed that the overconfidence effect is not as

general as previously thought (Juslin, Olsson, & Björkman, 1997; Juslin, Winman, & Olsson, 2000).

Another example is research on risky choices, where traditionally participants have been presented with summary descriptions of different options. Later research has shown that risky choices can be very different when people sample from the options themselves rather than relying on a description (Hertwig, Barron, Weber, & Erev, 2004; Lejarraga & Hertwig, 2021; Wulff, Mergenthaler-Canseco, & Hertwig, 2018). Relying solely on prior psychological studies to understand risky choice would not discover these insights.

Of course, new dimensions can always be added to the design space as they are discovered by new research, but this poses a practical problem of the rapidly growing number of experiments that could potentially be conducted. We therefore propose two ideas for a more exhaustive construction of the design space. One is to sample the phenomenon of interest directly. For example, Brunswik would sample participants' behavior in random intervals during several weeks, recording the behavior of interest as it occurs in the participants' natural environments (Brunswik, 1944). With today's technological developments, such experience-based sampling becomes easier to do and might be a way toward a more exhaustive grasp of the phenomenon of interest.

The other way to improve the construction of the design space is to do it collectively by many labs, in particular labs situated in different disciplines. For example, decades of research in social psychology suggest many different biases in human social cognition, which are often contradictory (Krueger & Funder, 2004). A tighter integration of psychology and network science has enabled recognizing how some of these biases in fact reflect a well-adapted cognition in specific social network structures (Dawes, 1989; Galesic, Olsson, & Rieskamp, 2018; Lee et al., 2019; Lerman, Yan, & Wu, 2016).

Second, how to deal with adaptive nature of complex social systems? As the authors point out, social and behavioral phenomena are typically caused by many interacting factors that can be hard to pin down. An additional, often overlooked property of these social-cognitive systems is that they are adaptive: They change over time in response to internal and external factors. As a consequence, even the most detailed static picture of these systems would not provide the full understanding of the underlying dynamics. This of course is a problem for both one-shot and integrative experiments, and it can be addressed by conducting longitudinal studies of these systems, coupled with theoretical development. For integrative experiments, however, it introduces the additional complication and cost of longitudinal studies, which multiplies the already large number of dimensions of the design space.

This explosion of potentially important dimensions in integrative experiment design could be tamed by assigning a stronger role to theory and modeling. The article focuses mostly on their role in interpreting the results of samples taken from an already constructed design space. However, theory and computational models seem essential already in the construction of the design space. In particular, an integrative theoretical framework constructed by a collective, strongly interdisciplinary effort mentioned above, could be a useful starting point for developing the initial design space. Such collective effort could also help recognize parts of the space that are implausible and would hardly be expected to occur in the real world. Then, computational modeling could be used to further narrow down the space by

investigating which of the dimensions could have a meaningful influence on the results. Such models could show that some apparently important dimensions have only a marginal influence on the system performance. Recognizing this could significantly narrow the otherwise vast space of possible experiments that could be run.








Third, what does it mean when results of experiments at particular points in the design space fail to generalize to other points? The authors suggest that this might point to an important missing dimension or even a fundamental limit of explanation of a particular phenomenon. It is however also possible that the reason is more prosaic, merely reflecting an inevitable random measurement error. This suggests that the integrative design experiments, just as one-at-a-time experiments, should be replicated. This would allow researchers to approximate confidence intervals around each of the samples from the design space and recognize what apparent differences between different points can be expected by chance. Moreover, it is likely that beyond random error, experiments conducted by any single lab will have some systematic biases stemming from lab-specific practices that can be hard to recognize without explicitly comparing labs. Different data analysts are also likely to reach different conclusions even from exactly the same data, so different labs conducting experiments from the same design space could reach different conclusions (Brenzau et al., 2022). To the extent that the integrative design experiments require resources that will limit them to a few larger labs, these biases could go unnoticed.

**Competing interest.** None.

## References

- Brenzau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., ... Van Assche, J. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, 119, e2203150119.
- Brunswik, E. (1944). Distal focussing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs*, 56, 1–49.
- Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25, 1–17.
- Galesic, M., Olsson, H., & Rieskamp, J. (2018). A sampling model of social judgment. *Psychological Review*, 125, 363–390.
- Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–528.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15, 534–539.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, 57, 226–246.
- Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the origin and nature of stochastic components of judgment. *Journal of Behavioral Decision Making*, 10, 189–209.
- Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, 107, 384–396.
- Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27, 313–327.
- Lee, E., Karimi, F., Wagner, C., Jo, H.-H., Strohmaier, M., & Galesic, M. (2019). Homophily and minority-group size explain perception biases in social networks. *Nature Human Behaviour*, 3, 1078–1087.
- Lejarraga, T., & Hertwig, R. (2021). How experimental methods shaped views on human competence and rationality. *Psychological Bulletin*, 147(6), 535–564.
- Lerman, K., Yan, X., & Wu, X. Z. (2016). The “majority illusion” in social networks. *PLoS ONE*, 11(2), e0147617.
- Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, 144, 140–176.

## Consensus meetings will outperform integrative experiments

Maximilian A. Primbs<sup>a</sup> , Leonie A. Dudda<sup>b,c</sup> ,  
Pia K. Andresen<sup>d</sup> , Erin M. Buchanan<sup>e</sup> , Hannah  
K. Peetz<sup>a</sup> , Miguel Silan<sup>f,g</sup>  and Daniël Lakens<sup>h\*</sup> 

<sup>a</sup>Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands;

<sup>b</sup>Department of Otorhinolaryngology, Head and Neck Surgery, University Medical Center, Utrecht, The Netherlands; <sup>c</sup>University Medical Center Utrecht Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands;

<sup>d</sup>Department of Methodology & Statistics, Utrecht University, Utrecht, The Netherlands; <sup>e</sup>Harrisburg University of Science and Technology, Harrisburg, PA, USA; <sup>f</sup>Anecy Behavioral Science Lab, Menthon Saint Bernard, France;

<sup>g</sup>Développement, individu, processus, handicap, éducation (DIPHE), Université Lumière Lyon 2, Bron Cedex, France and <sup>h</sup>Human–Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

[max.primbs@ru.nl](mailto:max.primbs@ru.nl), <https://max-primbs.netlify.app/>

[l.a.dudda@umcutrecht.nl](mailto:l.a.dudda@umcutrecht.nl)

[p.k.andresen@uu.nl](mailto:p.k.andresen@uu.nl)

[ebuchanan@harrisburgu.edu](mailto:ebuchanan@harrisburgu.edu), <https://www.aggieerin.com/>

[hannah.peatz@ru.nl](mailto:hannah.peatz@ru.nl)

[MiguelSilan@gmail.com](mailto:MiguelSilan@gmail.com)

[D.Lakens@tue.nl](mailto:D.Lakens@tue.nl), <https://sites.google.com/site/lakens2>

\*Corresponding author.

doi:10.1017/S0140525X23002248, e56

### Abstract

We expect that consensus meetings, where researchers come together to discuss their theoretical viewpoints, prioritize the factors they agree are important to study, standardize their measures, and determine a smallest effect size of interest, will prove to be a more efficient solution to the lack of coordination and integration of claims in science than integrative experiments.

Lack of coordination limits both the accumulation and integration of claims, as well as the efficient falsification of theories. How is the field to deal with this problem? We expect that consensus meetings (Fink, Kosecoff, Chassin, & Brook, 1984), where researchers come together to discuss their theoretical viewpoints, prioritize the factors they all agree are important to study, standardize their measures, and determine a smallest effect size of interest, will prove to be a more efficient solution to the lack of coordination and integration of claims in science than integrative experiments. We provide four reasons.

First, design spaces are simply an extension of the principles of multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) to theory-building. Researchers have recognized that any specified multiverse is just one of many possible multiverses (Primbs et al., 2022). The same is true for design spaces. People from different backgrounds and fields are aware of different literatures and might therefore construct different design spaces. Therefore, in practice a design space does not include all factors that members of a scientific community deem relevant – they merely include one possible subset of these factors. While any single design space can lead to findings that can be used to generate new hypotheses, it is not sufficient to *integrate* existing

hypotheses. Designing experiments that inform the integration of disparate findings requires that members of the community agree that the design space contains all relevant factors to corroborate or falsify their predictions. If any such factor is missing, members of the scientific community can more easily dismiss the conclusions of an integrative experiment for lacking a crucial moderator or including a condemning confound. Committing a priori to the outcome – for example, in a consensus meeting – makes it more difficult to dismiss the conclusions.

We believe that to guarantee that people from different backgrounds, fields, and convictions are involved in the creation and approval of the design space, consensus meetings will be required. During these consensus meetings, researchers will need to commit in advance to the consequences that the results of an integrative experiment will have for their hypotheses. Examples in the psychological literature show how initial versions of such consensus-based tests of predictions can efficiently falsify predictions (Vohs et al., 2021), and exclude competing hypotheses (Coles et al., 2022). Furthermore, because study-design decisions always predetermine the types of effects that can be identified in the design space, varying operationalizations may result in multiple versions of a study outcome that are not proforma comparable. To reduce the risks of a “methodological imperative” (Danziger, 1990), we need a consensus among experts on the theory and construct validity of the variables being tested.

Second, many of the observed effects in a partial design space will be either too small to be theoretically interesting, or too small to be practically important. Determining when effect sizes are too small to be theoretically or practically interesting can be challenging, yet it is essential to be able to falsify predictions, as well as to show the absence of differences between experiments (Primbs et al., 2023). Due to the combination of “crud” (Orben & Lakens, 2020) and large sample sizes, very small effect sizes could be statistically significant in integrative experiments. Without specifying a smallest effect of interest, the scientific literature will be polluted with a multitude of irrelevant and unfalsifiable claims. For integrative experiments, which require a large investment of time and money, discussions about which effects are large enough to matter should happen before data are collected. Many fields that have specified smallest effect sizes of interest have used consensus meetings to discuss this important topic.

Third, it is important to note that due to the large number of comparisons made in integrative experiments, some significant differences might not be due to crud (i.e., true effects caused by uninteresting mechanisms), but due to false positives. Strictly controlling the type 1 error rate when comparing many variations of studies will lower the statistical power of tests as the number of comparisons increases. Not controlling for multiple comparisons will require follow-up replication studies before claims can be made. Such is the cost of a fishing expedition. Consensus meetings, which have as one goal to reach collective agreement on which research questions should be prioritized, while coordinating measures and manipulations across studies, might end up being more efficient.

Fourth, *identifying* variation in effect sizes across a range of combinatorial factors is not sufficient to *explain* this variation. To make generalizable claims and distinguish hypothesized effects from confounding variables, one must understand how design choices affect effect sizes. Here, we consider machine-learning (ML) approaches a toothless tiger. Because these models exploit



all kinds of stochastic dependencies in the data, ML models are excellent at identifying predictors in nonexplanatory, predictive research (Hamaker, Mulder, & Van IJzendoorn, 2020; Shmueli, 2010). If there is a true causal model explaining the influence of a set of design choices and variables on a study outcome, the algorithm will find all relations – even those due to confounding, collider bias, or crud (Pearl, 1995). Algorithms identify predictors only relative to the variable set – the design space – so even “interpretable, mechanistic” (target article, sect. 3.3.1, para. 3) ML models cannot simply grant indulgence in causal reasoning. Achieving causal understanding through ML tools (e.g., through causal discovery algorithms) requires researchers to make strong assumptions and engage in a priori theorizing about causal dependencies (Glymour, Zhang, & Spirtes, 2019). Here again, we believe it would be more efficient to debate such considerations in consensus meetings.

We believe integrative experiments may be useful when data collection is cheap and the goal is to develop detailed models that predict variation in real-world factors. Such models are most useful when they aim to explain variation in naturally occurring combinations of factors (as effect sizes for combinations of experimental manipulations could quickly become nonsensical). For all other research questions where a lack of coordination causes inefficiencies, we hope researchers studying the same topic will come together in consensus meetings to coordinate their research.

**Competing interest.** None.

## References

- Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndakaihe, I. L. G., ... Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, 6(12), 1731–1742. <https://doi.org/10.1038/s41562-022-01458-9>
- Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511524059>
- Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: Characteristics and guidelines for use. *American Journal of Public Health*, 74(9), 979–983. <https://doi.org/10.2105/AJPH.74.9.979>
- Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 524. <https://doi.org/10.3389/fgene.2019.00524>
- Hamaker, E. L., Mulder, J. D., & Van IJzendoorn, M. H. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental Cognitive Neuroscience*, 46, 100867. <https://doi.org/10.1016/j.dcn.2020.100867>
- Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, 3(2), 238–247. <https://doi.org/10.1177/2515245920917961>
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688. <https://doi.org/10.2307/2337329>
- Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S. N., & Forscher, P. S., ... Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? A reply to Götz et al. (2022). *Perspectives on Psychological Science*, 18(2), 508–512. <https://doi.org/10.1177/17456916221100420>
- Primbs, M. A., Rinck, M., Holland, R., Knol, W., Nies, A., & Bijlstra, G. (2022). The effect of face masks on the stereotype effect in emotion perception. *Journal of Experimental Social Psychology*, 103, Article 104394. <https://doi.org/10.1016/j.jesp.2022.104394>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-sts330>
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., ... Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, 32(10), 1566–1581. <https://doi.org/10.1177/0956797621989733>

## Diversity of contributions is not efficient but is essential for science

Catherine T. Shea  and Anita Williams Woolley\* 

Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA  
[ctshea@andrew.cmu.edu](mailto:ctshea@andrew.cmu.edu)  
[awoolley@andrew.cmu.edu](mailto:awoolley@andrew.cmu.edu)  
<https://www.cmu.edu/tepper/faculty-and-research/faculty-by-area/profiles/shea-catherine.html>  
<https://scholars.cmu.edu/418-anita-woolley>

\*Corresponding author.

doi:10.1017/S0140525X23002285, e57

### Abstract

Dominant paradigms in science foster integration of research findings, but at what cost? Forcing convergence requires centralizing decision-making authority, and risks reducing the diversity of methods and contributors, both of which are essential for the breakthrough ideas that advance science.

The integrative experiment design approach advocated by Almaatouq et al. represents an intervention to accelerate the convergence of research findings in the social and behavioral sciences. Observations from the evolution of scientific fields over centuries lead to questions about whether the results of such an intervention would be uniformly positive. According to Kuhn (1962), all scientific fields go through initial periods, sometimes spanning centuries as in the case of physics, during which many concepts and competing models are proposed. This continues until a breakthrough insight reconciles discrepancies and establishes a dominant paradigm around which the field coheres. Dominant paradigms enable what Kuhn (1962) calls “normal science,” coordinated efforts to refine the paradigm and build evidence; however, the power structure surrounding a dominant paradigm can suppress alternative perspectives, making it difficult to prompt its reconsideration.

These observations from the history of science suggest potential unintended consequences of the intervention to accelerate convergence that Almaatouq et al. propose. Two sources of concern are the power structures that typically evolve to maintain organizing paradigms, and the potential they have to overly constrain the breadth of inputs considered, both of which can be problematic for a young science focused on diverse, multifaceted phenomena.

### Who decides?

Pfeffer (1993) cautioned that fields with higher levels of consensus get that way via a core group of elite scholars who wield control. Imposing a framework to foster consensus requires some mechanism for decision making. For instance, what variables are included or receive more attention? When two groups of researchers have converged on the same topic, who gets the naming rights to the theoretical space? These issues are often sorted out via peer-review processes and citation of papers, which admittedly is not “efficient” but incorporates the judgments of many other

researchers in the field, based on their assessment of the evidence. And, contrary to the authors' claim that no integrating frameworks exist, we point to a few recent examples in work on team process (i.e., Marks, Mathieu, & Zaccaro, 2001) and team structure (i.e., Hollenbeck, Beersma, & Schouten, 2012) that have been built upon by others based on the evidence supporting them. In the approach proposed by Almaatouq et al., the dimensions are "mapped" onto the design space *before* the experiment is run, by a "cartographer" – but how does this occur?

The solution – as proposed – is for machine learning to make such decisions for us. While an elegant solution sidestepping the potential of an individual or group of individuals making the decisions, machine-learning algorithms by their very nature have bias baked into them (e.g., Fu, Aseri, Singh, & Srinivasan, 2022). Furthermore, while machine learning undoubtedly has a substantial role to play in many areas of science, machine-learning models can only analyze the information provided, and are typically not able to identify variables that have not yet been considered but should be. While Almaatouq et al. would argue that such machines are flexible, and adaptive to change, we see this as overly optimistic. Indeed, across multiple disciplines and experiments, we can say with some certainty that a status quo – once set – is very difficult to change, as Kuhn's (1962) observations of the difficulty of challenging "dominant paradigms" demonstrates.

Another significant challenge stems from the strong incentives for researchers to introduce novel ideas in order to advance their careers. As Almaatouq et al. acknowledge, these incentives are at odds with efforts to promote convergence since there are few rewards for researchers who contribute to "normal science." Though the authors attempt to brush this aside by pointing to examples from physics, it is important to note that fields requiring major infrastructure investment also tend to be more hierarchical, and also have significant struggles with other issues such as sexual harassment as well as gender gaps in participation and career length (Huang, Gates, Sinatra, & Barabási, 2020; National Academies of Science, Engineering, and Medicine, 2018). Thus the efficiency that can come from centralizing decision-making authority to accelerate convergence also risks introducing some of the known problems associated with consolidating power (Pfeffer, 1993).

### Limiting diversity

The imposition of a framework for fostering convergence not only risks creating problematic power dynamics but also limiting the diversity of ideas in undesirable ways. Almaatouq et al. argue that their framework enables different studies to make their measures "commensurable." They claim this can facilitate the integration of research using different methods, however, it most naturally lends itself to the use of the "high-throughput" techniques they mention, typically online experiments, to generate the volume of data needed for sampling the design space. This is likely to result in more uniformity in the methods and measures used. While some may see this as desirable, we point out that when different studies using different methods yield convergent patterns of results, the field can have greater confidence in those effects, as advocates of "full-cycle research" (e.g., Chatman & Flynn, 2005) point out. Conversely, findings using the same measures and methods might make the results of different studies "commensurable," but could mask limits to generalizability. Indeed, just as we have made great strides to sample beyond undergraduate students, we need to continue to push scientists to replicate and extend their work beyond that of online samples,

as such samples are limited in their ability to capture rich behavioral outcomes. We also need to continue to broaden connections across contexts and disciplines to enable surprising new breakthroughs to emerge (Shi & Evans, 2023).

The social and behavioral sciences are at an exciting nexus. Diversity is finally gaining traction: Historically underrepresented groups are bringing new theory and ideas to our historically homogeneous field. Taken-for-granted knowledge is being falsified, or shown to only apply to the dominant groups. Exciting perspectives are just now being brought to fruition. To borrow from the author's terminology, the "unknown unknowns" are just starting to emerge due to the burgeoning diversity in the field. Will kicking off an intervention to force convergence, facilitated by machine learning, bake today's bias into algorithms that stymie the diversity that is just starting to take hold in our fields (Daft & Lewin, 1990)?

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

### References

- Chatman, J. A., & Flynn, F. J. (2005). Full-cycle micro-organizational behavior research. *Organization Science*, 16(4), 434–447.
- Daft, R. L., & Lewin, A. Y. (1990). Can organization studies begin to break out of the normal science straitjacket? An editorial essay. *Organization Science*, 1(1), 1–9.
- Fu, R., Aseri, M., Singh, P. V., & Srinivasan, K. (2022). "Un"fair machine learning algorithms. *Management Science*, 68(6), 4173–4195. <https://doi.org/10.1287/mnsc.2021.4065>
- Hollenbeck, J. R., Beersma, B., & Schouten, M. E. (2012). Beyond team types and taxonomies: A dimensional scaling conceptualization for team description. *Academy of Management Review*, 37(1), 82–106.
- Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 4609–4616. <https://doi.org/10.1073/pnas.1914221117>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26(3), 356–376.
- National Academies of Sciences, Engineering, and Medicine. (2018). *Sexual harassment of women: Climate, culture, and consequences in academic sciences, engineering, and medicine*. National Academies Press. <https://doi.org/10.17226/24994>
- Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review*, 18(4), 599–620.
- Shi, F., & Evans, J. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1), Article 1. <https://doi.org/10.1038/s41467-023-36741-4>

## Phenomena complexity, disciplinary consensus, and experimental versus correlational research in psychological science

Dean Keith Simonton\* 

Department of Psychology, University of California, Davis, Davis, CA, USA  
[dksimonton@ucdavis.edu](mailto:dksimonton@ucdavis.edu)  
<https://simonton.faculty.ucdavis.edu/>

\*Corresponding author.

doi:10.1017/S0140525X23002339, e58

### Abstract

The target article ignores the crucial role of correlational methods in the behavioral and social sciences. Yet such methods are often mandated by the greater complexity of the phenomena investigated. This necessity is especially conspicuous in psychological research where its position in the hierarchy of the sciences implies the need for both experimental and correlational investigations, each featuring distinct assets.

Almaatouq et al. describe an innovative way to improve experimental research in the behavioral and social sciences. Yet one serious oversight in their proposed solution stands out: The complete omission of any discussion of correlational methods. Correlations are nowhere mentioned in the text nor is there any reference to the common statistical procedures associated with correlational research, such as multiple regression, factor analysis, and structural equation models. Correlational research is especially common in various social sciences, like sociology, cultural anthropology, political science, and economics, and such research plays a major role in psychological science as well. The last point was treated in a classic paper by Cronbach (1957) titled “The Two Disciplines of Scientific Psychology,” the two disciplines being experimental and correlational (see also Tracy, Robins, & Sherman, 2009). This bifurcation dates back to the earliest years of psychological research. Where Wilhelm Wundt founded experimental psychology, Francis Galton initiated correlational psychology, both in the latter half of the nineteenth century. But why do behavioral and social scientists adopt correlational methods when everybody knows that experimental methods are superior at making causal inferences? After all, “correlation can’t prove causation” has become a proverb in research methods courses.

Ironically, Almaatouq et al. themselves provide a partial answer when they note “Social and behavioral phenomena exhibit higher ‘causal density’ (or what Meehl called the ‘crud factor’) than physical phenomena, such that the number of potential causes of variation in any outcome is much larger than in physics and the interactions among these causes are often consequential” (target article, sect. 2.1, para. 2). In other words, physical phenomena are less complex than behavioral and social phenomena. To provide an illustration, when Newton formulated his universal law of gravity with respect to two bodies, he needed only three independent variables: The mass of each body and the distance between the body centers. With that key formula he could accurately predict both the trajectories of projectiles and the orbits of the known planets. In contrast, imagine what is necessary to account for the romantic attraction between two human bodies. Easily dozens of variables would be required – far more than the 20 in the target article title. These would include numerous demographic variables, personality traits, situational factors, and various determinants of physical attractiveness. Moreover, these variables would have to be combined in such a way as to allow for unrequited love, when one body is attracted but the other repelled, which has no counterpart in the physical world. That much given, it is extremely doubtful that even the most complicated equation would ever predict romantic attraction as precisely and universally as Newton’s gravitational formula. Many intangible factors, such as interpersonal “chemistry,” would necessarily be left out.

To be sure, phenomena complexity is by no means the only reason why researchers will adopt correlational rather than experimental methods. Experimenters must often face severe practical and ethical limitations that undermine their capacity to address certain significant questions. Variable manipulation and random assignment to treatment and control conditions are frequently rendered impossible except under draconian circumstances (as in Nazi death camps). Nevertheless, in this commentary I would like to focus on the complexity issue because that relates most closely to the rationale for the integrative experiment design advocated by Almaatouq et al. (for other important implications, see Sanbonmatsu, Cooley, & Butner, 2021; Sanbonmatsu & Johnston, 2019).

The philosopher August Comte (1839–1842/1855) was the first to suggest that the empirical sciences – those that deal with concrete subject matter (unlike abstract mathematics) – can be arrayed into a hierarchy. One of the criteria that he used to determine a discipline’s ordinal placement was the complexity of the phenomena that are the target of investigation. This complexity helped explain why certain sciences, such as astronomy, were able to emerge and mature prior to other sciences, such as biology. Because the sciences were not defined in the same way in Comte’s time, the social and behavioral sciences being largely absent, his scheme has undergone some modifications to make it more consistent with modern disciplinary categories (Cole, 1983). This transformation then supported empirical research on whether Comte’s hierarchy of the sciences could claim any validity (Benjafeld, 2020; Fanelli, 2010; Fanelli & Glänzel, 2013; Simonton, 2004, 2015; Smith, Best, Stubbs, Johnston, & Archibald, 2000). The hierarchy has been validated using multiple indicators, almost entirely objective but also including subjective ratings of the relative “hardness” of disciplines. The following results are representative (Simonton, 2015):

Physics > Chemistry >> Biology > Psychology >>> Sociology

Here the physical sciences come first, followed by biology and psychology, and then sociology. The number of “>” symbols indicates the degree of separation, for the hierarchy is quantitative, not just ordinal. Thus, physics and chemistry are very close, and so are biology and psychology, but biology is more distant from chemistry and sociology is even more distant from psychology.

It is noteworthy that this hierarchy strongly corresponds with disciplinary consensus about what constitute the key findings in the field (Simonton, 2015), the very problem that Almaatouq et al. were trying to solve with their integrative experiment design. Yet it is evident that the hierarchy also aligns inversely with the relative prominence of experimental versus correlational methods, with psychology appropriately placed near the middle. However, that should not lead us to belittle correlational research as inferior. In fact, such methods feature definite advantages over the experimental. Probably the most significant is the multivariate response to the 20 questions problem: Why not answer all 20 questions at once using the suitable number of independent variables? And why not use multiple indicators that simultaneously implement alternative operational definitions for the central substantive variables? In short, why not use latent-variable structural equation models? It is not psychology or sociology’s fault that their phenomena are often so complex that these



models will incorporate many more variables than a typical experiment. Better yet, within psychological science, those subdisciplines that are more correlational exhibit higher replication rates than those that are more experimental (Youyou, Yang, & Uzzi, 2023; see, e.g., Soto, 2019). That advantage certainly deserves consideration.

**Acknowledgment.** I thank Richard W. Robins for help in identifying two essential references.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Benjafield, J. G. (2020). Vocabulary sharing among subjects belonging to the hierarchy of sciences. *Scientometrics*, 125, 1965–1982. <https://doi.org/10.1007/s11192-020-03671-7>
- Cole, S. (1983). The hierarchy of the sciences? *American Journal of Sociology*, 89, 111–139. <https://doi.org/10.1086/227835>
- Comte, A. (1839–1842/1855). *The positive philosophy of Auguste Comte* (H. Martineau, Trans.). New York: Blanchard. (Original work published 1839–1842).
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–684. <https://doi.org/10.1037/h0043943>
- Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PLoS ONE* 5(4), e10068. doi:10.1371/journal.pone.0010068
- Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLoS ONE*, 8(6), e66938. doi:10.1371/journal.pone.0066938
- Sanbonmatsu, D. M., Cooley, E. H., & Butner, J. E. (2021). The impact of complexity on methods and findings in psychological science. *Frontiers in Psychology*, 11, 580111. <https://doi.org/10.3389/fpsyg.2020.580111>
- Sanbonmatsu, D. M., & Johnston, W. A. (2019). Redefining science: The impact of complexity on theory development in social and behavioral research. *Perspectives on Psychological Science*, 14, 672–690. <https://doi.org/10.1177/1745691619848688>
- Simonton, D. K. (2004). Psychology’s status as a scientific discipline: Its empirical placement within an implicit hierarchy of the sciences. *Review of General Psychology*, 8, 59–67. <https://doi.org/10.1037/1089-2680.8.1.59>
- Simonton, D. K. (2015). Psychology as a science within Comte’s hypothesized hierarchy: Empirical investigations and conceptual implications. *Review of General Psychology*, 19, 334–344. <https://doi.org/10.1037/gpr0000039>
- Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences. *Social Studies of Science*, 30, 73–94. <https://doi.org/10.1177/030631200030001003>
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, 30, 711–727. <https://doi.org/10.1177/0956797619831612>
- Tracy, J. L., Robins, R. W., & Sherman, J. W. (2009). The practice of psychological science: Searching for Cronbach’s two streams in social-personality psychology. *Journal of Personality and Social Psychology*, 96, 1206–1225. <https://doi.org/10.1037/a0015173>
- Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6), e2208863120. <https://doi.org/10.1073/pnas.2208863120>

## The miss of the framework

Paul E. Smaldino<sup>a,b\*</sup> 

<sup>a</sup>Department of Cognitive and Information Sciences, University of California, Merced, Merced, CA, USA and <sup>b</sup>Santa Fe Institute, Santa Fe, NM, USA  
psmaldino@ucmerced.edu  
<https://smaldino.com/>

\*Corresponding author.

doi:10.1017/S0140525X23002315, e59

## Abstract

The authors rightly critique existing social sciences approaches. However, they are too quick to dismiss the criticism that their proposed paradigm is atheoretical. Social and cognitive theories are indeed incommensurate, often due to the lack of a unifying framework. Without proper integration with theoretical frameworks, their proposal may merely produce a resource-intensive veneer of thoroughness without substantive improvements to understanding.

The authors have produced a valuable and timely critique of widespread approaches to social science research, and I found much to agree with in their essay. I agree with their claim that many problems in science are not solved by replicability, nor by any methods that improve the reliability of experiments (though these measures are still valuable, as reliability of results is a necessary but insufficient condition for robust science). I agree that experiments must be better integrated with theory, and that the cumulative advance of theoretical explanations is a fundamental goal of science (even if other goals can also exist simultaneously). And I agree that coherence across results and experiments is critical, and troublingly lacking in much of the social sciences (Smaldino, 2019). Nevertheless, I find their approach to theory development to be a bit hasty.

The authors toe a messy line in their critique of the “one-at-a-time” approach. Of course, all scientific explanations leave out large swaths of the complexity of real life. As von Uexküll (1921) noted over a century ago, it is only by doing “violence to reality” that science is possible. All scientific theories decompose their target systems into an artificial set of parts, properties and relationships. The trouble, in my view, comes not from trying to construct theories about social systems, but from overconfidence that the particular decomposition associated with a particular theory constitutes a satisfactory explanation of phenomena.

For the purpose of further elaboration, allow me to propose a distinction between hypotheses, theories, and theoretical frameworks (taken from Smaldino, 2023). A *hypothesis* is a prediction that if a particular set of assumptions are met, a particular set of consequences will follow. It is easy to see how problems arise if hypotheses are tested in isolation. A *theory* is a set of assumptions upon which hypotheses derived from that theory must depend. Strong theories allow us to generate clear and falsifiable hypotheses. However, different theories may decompose reality in different ways and may address qualitatively different questions about a particular system, making comparisons of competing theories challenging. A *theoretical framework* is a broad collection of related theories that all share a common set of core assumptions. An example of a theoretical framework is Darwinian evolution by natural selection, from which many subordinate theories have been derived. A robust framework provides the conditions for the accumulation of scientific understanding, because consistency between related theories must be constantly assessed. I think it is fair to say that there is not currently a single dominant framework for the social sciences. One likely reason is that there have been few incentives to develop one. Indeed, there may have been active selection *against* proclivities to do so, as that pursuit rarely leads to easily measured success in the increasingly cutthroat game of academic science. A single framework may also be undesirable, as it may preclude useful decompositions needed for certain theories (*contra* Popper, 1994).

The “integrative” approach proposed by the authors falls short in its overreliance on data and its dismissal of the importance of mechanistic or generative explanations. The approach provided *does* try to draw consistency across experiments, and this is laudable. But it underplays the value of consistent theoretical framework. This is demonstrated most clearly by the authors’ implication that an interpretable, mechanistic model is essentially equivalent to a “surrogate model,” which is able to generate data that look like those collected empirically while remaining agnostic to similarities in the data-generating processes. I find this implication troubling. One reason is aesthetic – it is more satisfying to have a realistic explanation for a process than to simply produce an alternative process that generates similar outcomes. If the only objection were aesthetic, it would be easy to dismiss as mere preference. But the distinction is actually much more serious than this. A model that accurately represents the mechanisms that generate data is necessarily robust to changes in the contextual conditions under which the data are generated. This is because the assumptions of the model accurately map onto the conditions of the real world (within reason – all maps are ultimately imprecise). So the model can therefore be adjusted to match the new conditions, or at least will help us to identify the data needed to revise the model to match those conditions. A surrogate model, on the contrary, cannot do this, because the mapping between the model assumptions and the real world is fundamentally inaccurate. Consider how financial models failed to predict the economic crash of 2008, because their models were not mechanistic and therefore relied on correlations which suddenly failed to hold (this was not their only failure).

Thankfully, there are already theoretical frameworks that underpin some robust, testable, and coherent theories of human behavior. These include cultural evolution (Boyd & Richerson, 1985; Cavalli-Sforza & Feldman, 1981; Mesoudi, 2011) and human behavioral ecology (Nettle, Gibson, Lawson, & Sear, 2013; Smith & Winterhalder, 1992), which draw on insights from biological theories of evolution and ecology, as well as from related work in microeconomics and game theory. These frameworks give us good prior reasons for incorporating certain assumptions into our theories while excluding others, because they relate to fundamental aspects of social life, such as the presence or absence of particular social learning biases or the use of prosocial norms as mechanisms for dealing with uncertainty and risk. One advantage of these frameworks is that they do not discard, as many other approaches do, the troves of knowledge we have acquired about nonhuman species. Since humans are, after all, also animals, we are subject to many similar constraints and affordances that occur in other species. Consistency with *these* data is too often overlooked.

**Competing interest.** None.

## References






- Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.
- Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.
- Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press.
- Nettle, D., Gibson, M. A., Lawson, D. W., & Sear, R. (2013). Human behavioral ecology: Current research and future prospects. *Behavioral Ecology*, 24(5), 1031–1040.
- Popper, K. R. (1994). *The myth of the framework: In defence of science and rationality*. Routledge.
- Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, 575, 9.

Smaldino, P. E. (2023). *Modeling social behavior: Mathematical and agent-based models of social dynamics and cultural evolution*. Princeton University Press.

Smith, E. A., & Winterhalder, B. (1992). *Evolutionary ecology and human behavior*. Routledge.

von Uexküll, J. (1921). *Umwelt und innenwelt der tiere*. Springer-Verlag.

## Are language–cognition interactions bigger than a breadbox? Integrative modeling and design space thinking temper simplistic questions about causally dense phenomena

Debra Titone<sup>a,b,c\*</sup> , Esteban Hernández-Rivera<sup>a,b,c</sup> , Antonio Iniesta<sup>a,b,c</sup> , Anne L. Beatty-Martínez<sup>a,b,c,d</sup>  and Jason W. Gullifer<sup>e</sup> 

<sup>a</sup>Department of Psychology, McGill University, Montreal, QC, Canada; <sup>b</sup>Montreal Bilingualism Initiative, Montreal, QC, Canada; <sup>c</sup>Centre for Research on Brain, Language, and Music, McGill University, Montreal, QC, Canada; <sup>d</sup>Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA and <sup>e</sup>Department of Computer Science, Marianopolis College, Westmount, QC, Canada

[debra.titone@mcgill.ca](mailto:debra.titone@mcgill.ca)

[esteban.hernandezrivera@mcgill.ca](mailto:esteban.hernandezrivera@mcgill.ca)

[antonio.martineziniesta@mcgill.ca](mailto:antonio.martineziniesta@mcgill.ca)

[abeattymartinez@ucsd.edu](mailto:abeattymartinez@ucsd.edu)

[j.gullifer@maranopolis.edu](mailto:j.gullifer@maranopolis.edu)

\*Corresponding author.

doi:10.1017/S0140525X23002145, e60

### Abstract

We affirm the utility of integrative modeling, according to which it is advantageous to move beyond “one-at-a-time binary paradigms” through studies that position themselves within realistic multidimensional design spaces. We extend the integrative modeling approach to a target domain with which we are familiar, the consequences of bilingualism on mind and brain, often referred to as the “bilingual advantage.” In doing so, we highlight work from our group consistent with integrative modeling.

The history of science abounds with self-reflections about whether its questions, methods, and theories are sufficiently rigorous to clarify complex unknowns. Metascientific accounts pervade our own fields of language and cognition, which coincidentally coalesced when “20 Questions” was a popular television show (Van Deventer, 1952). After Newell’s (1973) prescient warnings about playing 20 questions with Nature, current views about language–cognition interactions vary along many metascientific dimensions. We are thus grateful to Almaatouq et al. for reanimating Newell’s proposal in their paper, which names, operationally defines, and advocates for an *integrative modeling approach*.

Cognitive scientists have long debated how language and cognition interact. These debates take many forms, including the consequences of bilingualism on mind and brain, often referred

to as the “bilingual advantage” (see Titone, Gullifer, Subramanipillai, Rajah, & Baum, 2017, for historic overview). The initial rationale of this hypothesis is that people who speak multiple languages have heightened daily experience suppressing/inhibiting knowledge of one language when speaking another (e.g., Bialystok, Craik, Klein, & Viswanathan, 2004). Because researchers presumed that suppression/inhibition is part of a domain-general cognitive control capacity, the bilingual advantages position hypothesized that this daily practice would preferentially strengthen cognitive control for bilinguals compared to monolinguals, causing them to perform better on cognitive control tasks.

When the bilingual advantage hypothesis emerged (Bialystok et al., 2004; see also Peal & Lambert, 1962), it was refreshing in its celebration of bilinguals’ cognitive capacities compared to biased and culturally damaging notions of bilingualism as a liability (e.g., Goodenough, 1926; Saer, 1923). Nevertheless, it was much too simple in a “20-questions,” yes–no binary way. While early findings were supportive, it did not take long for mixed findings to emerge. Relevant to our commentary are researchers’ attributions for the sources of these mixed findings, which we class in two nonmutually exclusive ways – a “replication crisis” account, and – building upon Almaatouq et al. – an “integrative modeling/design space” account.

A “replication crisis” account presumes that replicable findings are true, and nonreplicable findings are false. However, jumping to conclusions prematurely risks perpetuating a 20-questions mindset by presuming that all studies are interchangeable (i.e., commensurate), when they may differ in a myriad of incommensurate ways (e.g., Are bilingual and monolingual groups comparably designated? Are all bilinguals the same in terms of language and cognitive experiences? Are all geographies equally supportive of bilingualism? Are all cognitive tasks equivalent? Does suppression/inhibition mean the same thing across all cognitive tasks?). Further, a potentially erroneous corollary of a reflexive replication crisis view is that there is one general cognitive reality applicable to all bilingual people, and that any experiment is an equipotent reflection of that reality.

In contrast, an “integrative modeling/design space” account takes mixed findings at face value and actively accounts for systematic differences across study details that could have elicited them. Indeed, much of our field has moved into this post-20-questions phase of inquiry (e.g., Navarro-Torres, Beatty-Martínez, Kroll, & Green, 2021), and now investigates the links between individual differences among bilinguals and a variety of performance outcomes (e.g., Wagner, Bekas, & Bialystok, 2023). As one example, our group developed new tools and methods for capturing nuanced differences among bilinguals (language entropy, social network analysis), including analytic approaches (e.g., machine-learning approaches such as leave one out cross-validation) that distinguish explanation and prediction, referred to in the target article (Gullifer, Pivneva, Whitford, Sheikh, & Titone, 2023; Gullifer & Titone, 2021; see also Hofman et al., 2021).

As another example compatible with the target article’s research cartography idea, our group posited the *systems framework of bilingualism* (Titone & Tiv, 2023; Tiv, Gullifer, Feng, & Titone, 2020; Tiv et al., 2022; see also Beatty-Martínez & Titone, 2021), which sketches a design space for language–cognition interactions. This framework builds upon socioecological accounts of human behavior (e.g., Atkinson et al., 2016; Bronfenbrenner, 1977; de Bot, Lowie, & Verspoor, 2007), and our prior efforts to encourage researchers to abandon simple bilingual/monolingual group comparisons for tasks that may not tap into the same cognitive

constructs (e.g., Baum & Titone, 2014; Beatty-Martínez & Titone, 2021, 2024; Gullifer & Titone, 2021; Titone & Baum, 2014; Tiv et al., 2020). Accordingly, people’s individual language and cognitive behaviors are embedded within a multilevel set of nested social influences (i.e., daily interactions, local neighborhoods, laws regulating language use). Thus, to fully describe language–cognition interactions among bilinguals (or anyone), one must attend to these influences, and how participants across studies systematically vary in these ways. This means that any one study is but a single point within a much larger space, that mixed findings may be meaningful, and that conclusions about bilingualism may be less general or unitary than one might originally believe. Such an approach respects the complexity of the phenomena such that, regardless of where the data ultimately lead, our conclusions will be more rigorously and honestly earned.

In closing, we agree with Almaatouq et al. that it is advantageous to move beyond “one-at-a-time binary paradigms” through studies that position themselves within realistic multidimensional design spaces (i.e., a preplanned meta-analytic approach). We are ever mindful that our work on language and cognition is conducted within a unique multilingual city where language use is legally regulated and often interpersonally, culturally, and politically charged. Consequently, what is possible for us to capture empirically about language–cognition interactions will be necessarily impacted by our unique positionality. Importantly, we are not alone, as every research group has its own unique positionality that must be considered. Thus, let us profit from the wisdom and humility implicit in the Almaatouq et al.’s target article and Newell’s (1973) original proposal, by recognizing that it may not be possible for any one experiment or research group to speak definitively to an entire design space of causally dense, socially situated behavioral phenomena.

**Financial support.** The authors gratefully acknowledge past and current support from the following sources: Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (Titone), Canada Research Chairs (Titone), the Social Sciences and Humanities Research Council Insight and Insight Development Grants (Titone, Gullifer), National Council for Science and Technology – CONACyT (Hernández-Rivera), and National Institutes of Health (F32-AG064810 to Beatty-Martínez and Gullifer).

**Competing interest.** None.

## References

- Atkinson, D., Byrnes, H., Doran, M., Duff, P., Ellis, N. C., Hall, J. K., ... Tarone, E. (2016). A transdisciplinary framework for SLA in a multilingual world. *Modern Language Journal, 100*, 19–47. <https://doi.org/10.1111/modl.12301>
- Baum, S., & Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics, 35*(5), 857–894. <https://doi.org/10.1017/S0142716414000174>
- Beatty-Martínez, A. L., & Titone, D. A. (2021). The quest for signals in noise: Leveraging experiential variation to identify bilingual phenotypes. *Languages, 6*(4), 168. <https://doi.org/10.3390/languages6040168>
- Beatty-Martínez, A. L., & Titone, D. A. (2024). De-generacy as an organizing principle of bilingual language processing: Evidence from brain and behavior. In K. Morgan Short & J. G. van Hell (Eds.), *The Routledge handbook of second language acquisition and neurolinguistics* (p. 260–273). Routledge.
- Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging, 19*(2), 290–303. <https://doi.org/10.1037/0882-7974.19.2.290>
- Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513–531. <https://doi.org/10.1037/0003-066X.32.7.513>
- de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition, 10*(1), 7–21. <https://doi.org/10.1017/S1366728906002732>
- Goodenough, F. L. (1926). Racial differences in the intelligence of school children. *Journal of Experimental Psychology, 9*(5), 388–397. <https://doi.org/10.1037/h0073325>



- Gullifer, J. W., Pivneva, I., Whitford, V., Sheikh, N. A., & Titone, D. (2023). Bilingual language experience and its effect on conflict adaptation in reactive inhibitory control tasks. *Psychological Science*, 34(2), 238–251. <https://doi.org/10.1177/09567976221113764>
- Gullifer, J. W., & Titone, D. (2021). Engaging proactive control: Influences of diverse language experiences using insights from machine learning. *Journal of Experimental Psychology: General*, 150(3), 414–430. <https://doi.org/10.1037/xge0000933>
- Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181–188. <https://doi.org/10.1038/s41586-021-03659-0>
- Navarro-Torres, C. A., Beatty-Martínez, A. L., Kroll, J. F., & Green, D. W. (2021). Research on bilingualism as discovery science. *Brain and Language*, 222, 105014. <https://doi.org/10.1016/j.bandl.2021.105014>
- Newell, A. (1973). *You can't play 20 questions with nature and win: Projective comments on the papers of this symposium*. <https://doi.org/10.1016/B978-0-12-170150-5.50012-3>
- Peal, E., & Lambert, W. E. (1962). The relation of bilingualism to intelligence. *Psychological Monographs: General and Applied*, 76(27), 1–23. <https://doi.org/10.1037/h0093840>
- Saer, D. J. (1923). The effect of bilingualism on intelligence. *British Journal of Psychology*, 14, 25–38.
- Titone, D., & Baum, S. (2014). The future of bilingualism research: Insufferably optimistic and replete with new questions. *Applied Psycholinguistics*, 35(5), 933–942. <https://doi.org/10.1017/S0142716414000289>
- Titone, D., Gullifer, J., Subramaniapillai, S., Rajah, N., & Baum, S. (2017). Chapter 13. History-inspired reflections on the bilingual advantages hypothesis. In E. Bialystok & M. D. Sullivan (Eds.), *Growing old with two languages: Effects of bilingualism on cognitive aging* (pp. 265–295). John Benjamins Publishing Company. <https://doi.org/10.1075/sibil.53.13tit>
- Titone, D. A., & Tiv, M. (2023). Rethinking multilingual experience through a systems framework of bilingualism. *Bilingualism: Language and Cognition*, 26(1), 1–16. <https://doi.org/10.1017/S1366728921001127>
- Tiv, M., Gullifer, J. W., Feng, R. Y., & Titone, D. (2020). Using network science to map what Montréal bilinguals talk about across languages and communicative contexts. *Journal of Neurolinguistics*, 56, 100913. <https://doi.org/10.1016/j.jneuroling.2020.100913>
- Tiv, M., Kutlu, E., Gullifer, J. W., Feng, R. Y., Doucerain, M. M., & Titone, D. A. (2022). Bridging interpersonal and ecological dynamics of cognition through a systems framework of bilingualism. *Journal of Experimental Psychology: General*, 151(9), 2128–2143. <https://doi.org/10.1037/xge0001174>
- Van Deventer, F. (1952). Twenty questions [television broadcast]. DuMont Television Network. <http://archive.org/details/TwentyQuestionsJanuary181952>
- Wagner, D., Bekas, K., & Bialystok, E. (2023). Does language entropy shape cognitive performance? A tale of two cities. *Bilingualism: Language and Cognition*, 1–11. <https://doi.org/10.1017/S1366728923000202>

## Experiment commensurability does not necessitate research consolidation

Milena Tsvetkova\* 

London School of Economics and Political Science, London, UK  
[m.tsvetkova@lse.ac.uk](mailto:m.tsvetkova@lse.ac.uk)  
<http://tsvetkova.me>

\*Corresponding author.

doi:10.1017/S0140525X23002364, e61

### Abstract

Integrative experiment design promises to foster cumulative knowledge by changing how we design experiments, build theories, and conduct research. I support the push to increase commensurability across experimental research but raise several reservations regarding results-driven and large-team-based research. I argue that it is vital to preserve academic diversity and adversarial debate via independent efforts.

The proposed integrative experiment design approach consists of three steps: (1) Explicitly define the design space of the experiments in terms of features of the decision situation and the population sample, (2) systematically sample from that design space, and (3) build theories by quantifying the outcome heterogeneity over that space. This approach will guarantee commensurability between different experiments and findings and foster cumulative knowledge. The authors' concept of "research cartography" is brilliant – the idea is to itemize, standardize, categorize, and quantify the information that we typically and only partially reveal in the Methodology and Discussion sections of our research papers. The image of a Wikidata-style database containing all experimental (and in fact, any other) social and behavioral knowledge is incredibly appealing! The Cooperation Databank, for instance, offers a glimpse of how such a database could look like (Spadaro et al., 2022). Developing research cartography will help identify research gaps, established findings, and controversial problems. The approach will also aid the reuse and reanalysis of existing data to answer new research questions (Almaatouq, Rahimian, Burton, & Alhajri, 2022; Rand, Greene, & Nowak, 2012; Tsvetkova, Wagner, & Mao, 2018). In short, whether retrospective or prospective, a comprehensive and systematic research cartography will help consolidate knowledge and stimulate new research.

The integrative experiment design approach, however, presses further – steps (2) and (3) propose to consolidate research and theory-testing efforts by sampling and generalizing over many points in the experimental design space simultaneously, rather than "one-at-a-time." Yet, these steps are not necessary for commensurability and more importantly, carry negative implications for diversity, innovation, and productive debate in academic research. There are several issues I would like to raise here.

First, the proposed paradigm threatens to entrench and exacerbate existing inequalities within and between scholarly communities. Participating in global research consortia may be open to many but who leads these consortia will likely befall on those with status, prestige, and funding. It is hard to overlook the fact that the authors speak from a position of privilege – they work at prestigious US universities, with access to hefty research funds and numerous PhD students and postdoctoral researchers. The large-scale research they propose is simply not accessible to many experimental researchers.

Second, the proposed paradigm aims to optimize efficiency in research but this is a misguided ideal. Academic research is not just about results but also about exploration and discovery, critique and debate, learning and training. Consolidating research activities in hierarchically structured labs or consortia with established protocols and routines may reduce labor costs but stifle entrepreneurship, critical thinking, and iconoclastic innovation. Based on some of the authors' empirical examples, the complexities of group synergies imply that different problems would be best addressed by teams of different size and composition (Almaatouq, Alsobay, Yin, & Watts, 2021; Mao, Mason, Suri, & Watts, 2016; Straub, Tsvetkova, & Yasseri, 2023). This calls for independence, plurality, redundancy, and diversification of research effort, not consolidation.

Third, in the social and behavioral sciences, raising the question is often more important than finding out the answer. Much like the observer effect related to measuring physical systems, studying a social system changes it. Posing a specific social research problem can shape political debate, policy decisions,

organization strategies, and collective behavior. Consolidating funding and research efforts forebodes a monopoly over setting research agendas and directions, the muffling of marginalized voices, the sidelining of localized problems, and the suppression of new perspectives and paradigms. Large-scale integrative experiments may be good for providing definitive evidence to integrate and reconcile existing theories but restricted when it comes to launching new research agendas.

Related to the latter issue, the proposed result-driven active-learning sampling strategy for experiments threatens to shift the focus to effects that are sizeable but not necessarily meaningful or important. Specifically, certain combinations of context and population features may be impossible or unlikely and hence, practically irrelevant. In short, the integrative experiment design approach does not alleviate and may even exacerbate the thorniest problem of experimental research – external validity. Explaining all variation is not always the best strategy for good or efficient science: The power of good general theories is not that they are universally true but that they apply to statistically likely/common situations and hence, they are useful.

I acknowledge that the authors present integrative experiment design as an additional, and not the only true, approach to experimentation in the social and behavioral sciences. I assumed an exaggeratedly antagonistic stance here to caution against consolidation. There are alternatives, such as adversarial collaboration (Killingsworth, Kahneman, & Mellers, 2023; Mellers, Hertwig, & Kahneman, 2001), that can help reconcile contradictory findings without compromising debate, plurality, and diversity.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences of the United States of America*, 118(36), e2101062118. <https://doi.org/10.1073/pnas.2101062118>
- Almaatouq, A., Rahimian, M. A., Burton, J. W., & Alhajri, A. (2022). The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Scientific Reports*, 12(1), Article 1. <https://doi.org/10.1038/s41598-022-20551-7>
- Killingsworth, M. A., Kahneman, D., & Mellers, B. (2023). Income and emotional well-being: A conflict resolved. *Proceedings of the National Academy of Sciences of the United States of America*, 120(10), e2208661120. <https://doi.org/10.1073/pnas.2208661120>
- Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PLoS ONE*, 11(4), e0153048. <https://doi.org/10.1371/journal.pone.0153048>
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12(4), 269–275. <https://doi.org/10.1111/1467-9280.00350>
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427–430. <https://doi.org/10.1038/nature11467>
- Spadaro, G., Tiddi, I., Columbus, S., Jin, S., ten Teije, A., CoDa Team, & Balliet, D. (2022). The Cooperation Databank: Machine-readable science accelerates research synthesis. *Perspectives on Psychological Science*, 17(5), 1472–1489. <https://doi.org/10.1177/17456916211053319>
- Straub, V. J., Tsvetkova, M., & Yasseri, T. (2023). The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence*, 2(2). <https://doi.org/10.1177/26339137231156912>
- Tsvetkova, M., Wagner, C., & Mao, A. (2018). The emergence of inequality in social groups: Network structure and institutions affect the distribution of earnings in cooperation games. *PLoS ONE*, 13(7), e0200965. <https://doi.org/10.1371/journal.pone.0200965>

## Eliminativist induction cannot be a solution to psychology's crisis

Mehmet Necip Tunç<sup>a\*</sup>  and Duygu Uygun Tunç<sup>b</sup> 

<sup>a</sup>Tilburg University, Tilburg, Netherlands and <sup>b</sup>Eindhoven University of Technology, Eindhoven, Netherlands  
[m.neciptunc@hotmail.com](mailto:m.neciptunc@hotmail.com)  
[duygu.uygun@outlook.com](mailto:duygu.uygun@outlook.com)

\*Corresponding author.

doi:10.1017/S0140525X23002157, e62

### Abstract

Integrative experiment design assumes that we can effectively design a space of factors that cause contextual variation. However, this is impossible to do so in a sufficiently objective way, resulting inevitably in observations laden with surrogate models. Consequently, integrative experiment design may even deepen the problem of incommensurability. In comparison, one-at-a-time approaches make much more tentative assumptions about the factors excluded from experiment design, hence still seem better suited to deal with incommensurability.

The authors address the problem of how to integrate the results of independent studies in a way that facilitates knowledge accumulation in psychological science. We agree with the authors that most experiments as they are currently conducted in psychology have low information value. The authors claim that this is because (1) in psychological science the phenomena to be explained are much more complex and the theories are not precise enough, so that theories cannot indicate which auxiliary assumptions might be safely relegated to the *ceteris paribus* clause, and (2) in the absence of precise enough theories the practice of designing experiments one-at-a-time hampers the goal of knowledge accumulation because the results of individual experiments are incommensurate. They infer from this diagnosis that reforming the scientific practices in psychology toward more reliable studies is misguided because however reliable individual studies are, they will nonetheless fail to fit together with one another in a way that enables knowledge accumulation. They propose that instead of increasing the reliability of individual studies, we should replace the one-at-a-time paradigm with integrative experiment design, which involves constructing a design space that defines all relevant contextual factors and then systematically testing their effects.

We underline two core problems with this proposal. The first is that the authors propose a solution that is fraught with intractable problems. The second is that the authors are mistaken in their diagnosis that incommensurability is an issue that only applies to hypothetico-deductive approaches that involve testing alternative explanations one-at-a-time.

The integrative experiment design strategy bears serious similarities to eliminativist induction, also known as the Baconian method. The essence of this method is that the researchers in a (sub)discipline first construct an event space in which the context variables are defined, and then, by eliminating alternative explanations they arrive at an inductive generalization. As long as the

defined event space effectively covers all aspects of the target phenomenon, the inductive inference made on the basis of observed instances will be accurate.

However, several philosophers of science such as Goodman (1983), Popper (1959), Quine (1951), and others have shown in various ways that this important assumption on which eliminativist induction is based is almost never true, that is, it is impossible to effectively map out the contextual variations of even a single phenomenon because the list has infinitely many elements. The only viable strategy, as the authors point out in line with what Bacon (1994) suggested four centuries ago, is to find the elements that make a significant difference. However, determining which factors would make a significant difference in the contextual variation space is an even harder problem in psychology, because, as the authors also admit, psychological phenomena are inherently more causally dense than natural science disciplines such as physics. But still, the authors suggest, again in line with the Baconian method, “conducting a small number of randomly selected experiments (i.e., points in the design space) and fitting a surrogate model” (target article, sect. 3.2, para. 31). However, since only an omniscient being can have the knowledge of a predefined contextual space, no such experiments can be truly random and hence the researchers cannot avoid the risk that their surrogate model overfits the experiments they perform. So, that the overfitted model would reflect the initial assumptions of the experimenters more than it reflects the underlying reality it purports to describe.

An active learning perspective that updates the surrogate model with new studies is not enough to solve this problem for two reasons. First, no matter how systematically we vary the experiments based on whose results we update our surrogate model, it is very likely that we will ignore critical contextual variables that the prevailing scientific paradigm of the time does not consider important (and thus not include them in the design space; also see Kuhn, 1977, for how paradigms shape even the basic observations). Second (and relatedly), the problem of weighing or appraising the novel evidence during the update always has to be surrogate model-laden. For example, because of the high heterogeneity pertaining to psychological phenomena, two alternative surrogate models with different sets of initial experiments will most probably incorporate different dimensions to be important, and thus might place the same experiment in radically different points in the design space. Even if these two surrogate models are attempted to be combined, which observations count as valid evidence and how these pieces of evidence are weighted will be a matter of debate among scientists advocating different surrogate models.

Consequently, (1) the main assumption of integrative experiment design is that one can effectively define a design space but it is an impossible task and (2) the problem of theory-ladenness and incommensurability will not be solved by integrative experiment design. Actually what the authors call “one-at-a-time approach” still has a better chance of addressing the incommensurability-related issues that arise from the inherent complexity of psychological phenomena, because it does not require researchers to commit themselves to any list of elements that causes contextual variation but, on the contrary, it requires the researchers to be actively on the search for contextual variables that behave in a way that is not predicted by their theory. So, it allows researchers to devise more severe tests to falsify their theory if it is indeed incorrect. Assuming that we can know at any point which elements of contextual variation are important is only

possible through an unjustified indifference to elements outside the design space we have already defined, and for this reason, methodologies that depend on this assumption can give us only an illusion of enabling knowledge accumulation about psychological phenomena. And since it would almost always be impossible to build a consensus among scientists with different perspectives about the elements that need to be in the design space, encountering the problem of incommensurability is also inevitable in integrative experiment design. Therefore, methods that depend on eliminativist induction, such as integrative experiment design, could not be an effective solution to psychology’s credibility crisis.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Bacon, F. (1994). *Novum organum*. Carus (original work published 1620).  
 Goodman, N. (1983). The new riddle of induction. In N. Goodman (Ed.), *Fact, fiction, and forecast* (pp. 59–83). Harvard University Press (original work published 1954).  
 Kuhn, T. S. (1977). Objectivity, value judgement, and theory choice. In T. S. Kuhn (Ed.), *The essential tension* (pp. 320–339). University of Chicago Press.  
 Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.  
 Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60, 20–43. doi:10.2307/2181906

## Commensurability engineering is first and foremost a theoretical exercise

Joachim Vandekerckhove\* 

University of California, Irvine, Irvine, CA, USA  
 joachim@uci.edu  
 www.cidlab.com

\*Corresponding author.

doi:10.1017/S0140525X23002224, e63

### Abstract

I provide a personal perspective on metastudies and emphasize lesser-known benefits. I stress the need for integrative theories to establish commensurability between experiments. I argue that mathematical social scientists should be engaged to develop integrative theories, and that likelihood functions provide a common mathematical framework across experiments. The development of quantitative theories promotes commensurability engineering on a larger scale.

When we first executed a metastudy in 2015 (Baribault, 2019; Baribault et al., 2018), the concept of sampling from a method space (what the target article calls a “design space”) was central to its implementation. We had set out to replicate an interesting effect we had found in a published paper. However, we soon realized that we would need to specify so many details of implementation – the kinds of things researchers rarely make explicit in their methods sections – that we felt we could not perform a faithful replication. Of course, we could have reached out to the original authors, but we also felt that the literature should to some



extent be able to stand on its own. Eventually, we decided to be good Bayesians and allow for uncertainty in our experimental design. In contrast to a “point experiment,” a metastudy defines a distribution over the method space, from which we can draw samples in a kind of Monte Carlo integration over our uncertainty as to which point experiment best captures the effect of interest.

Our intent was to test a particular type of theory: A statement that is broader than a single contrast or effect, but is about regions in the method space where an effect holds. Others have referred to such regions as the universe of generalization (Cronbach, Rajaratnam, & Gleser, 1963), constraints on generality (Simons, Shoda, & Lindsay, 2017), or the boundary of meaning (Kenett & Rubinstein, 2021) – all invoking metaphors that imply the existence of some spatially arranged population of possible experiments.

We were interested in exploring this method space in part to identify moderators of effects but also to establish invariances. Invariances were perhaps of greater interest because they speak to the robustness of effects across sets of *exchangeable experiments* – experiments that are not identical, but that are minor variations on each other such that a reasonable experimenter could have chosen any one of them to test the theory at hand. In other words, many randomly sampled experiments are *identical in theory*, if not necessarily so in practice. We focused on randomization specifically because we wanted to determine whether an effect was robust – that is, whether it was sensitive to irrelevant perturbations of the study, such as who the participants were, where the study was conducted, or which #@\$%&? masking symbol we chose.

This notion of *identity in theory* is important, I think. Whether two experiments can be reasonably compared or jointly analyzed (i.e., whether they are *commensurate*) depends not only on how they relate to one another but also on the theoretical weight given to that relationship. Without the context of germ theory, washing hands between patients may seem like a silly exercise, but in reality handwashing can act as an accidental confounder if it is not properly controlled. Accordingly, there must be a role for the formation of theories prior even to the construction of the method space.

The target article understates the importance of the development of integrative theory relative to the experimentation framework. Without a connecting theory, no two experiments (or, for that matter, observations) are commensurate. *With* a connecting theory, it does not seem to matter greatly if the method space was conceived ahead of time or even at all. Commensurability engineering – the activity of building experiments such that they are commensurate – is first and foremost a theoretical exercise. But this invites a new question: If indeed disparate experiments can be made commensurate with a properly integrative theory, and method spaces only provide commensurability if there is such a theory, then what justifies the added effort of designing a metastudy? After all, a space of experiments exists whether we define one or not and a research program of consecutive point experiments constitutes a guided walk in some space, so is not any collection of point experiments a metastudy?

An underappreciated strength of metastudies is their statistical efficiency (DeKay, Rubinchik, Li, & De Boeck, 2022; Rubinchik, 2019). In a metastudy, increasing the number of point experiments  $k$  reduces the standard error of the mean effect size above and beyond the total number of participants  $P$ . To see this, consider the equation for the error variance in a random-effects meta-analysis as a function of the variance in effect sizes

across subjects ( $\sigma^2$ ) and the variance in effect sizes across studies ( $\tau^2$ ):  $\sigma_{\delta}^2 = \sigma^2/P + \tau^2/k$ . For a fixed number of participants, increasing the number of point experiments (and reducing the number of participants per study) maximizes estimation accuracy.

Looking ahead, I believe there is much relevant work being done in the field of mathematical behavioral science. In order to engineer commensurability at scale, it is critical to develop *quantitative* integrated theories. Ideally these would take the form of likelihood functions – functions that describe the probability of data patterns under a theory – over the method space. A likelihood framework for theoretical integration has a number of advantages. For example, such a framework would be applicable even with complex theories for complex data. The focus of the target article seems mostly on linear theories – models that are composed mostly of effects (or “dependencies”) that change the mean of some variate in an additive or at most interactive way – but a well-constructed mathematical likelihood can account for patterns of any kind and data of any shape.

Even more importantly, likelihoods are inherently commensurate and can act as a universal language in which theories can be cast for comparison between areas of a method space (whether intentionally designed or not). Regions  $A$  and  $B$  of the method space are identical in theory  $T$  if they come with the same likelihood,  $p(\text{data} | A, T) = p(\text{data} | B, T)$ , and not otherwise. The development of an integrative theory then boils down to defining this likelihood for all applicable regions, making all points in the method space commensurate while at the same time avoiding the incoherency problem discussed by Watts (2017). Theories of such scope are currently rare in social science, but we stand to gain much from their development.

**Acknowledgment.** Unable to find a native English speaker for proofreading on short notice, I asked ChatGPT to evaluate my writing. It found my grammar and spelling to be “mostly on par” with a native English speaker, which I found comforting.





**Financial support.** J. V. was supported by NSF grant Nos. 1754205, 1850849, and 2051186.

**Competing interest.** None.

## References

- Baribault, B. (2019). Using hierarchical Bayesian models to test complex theories about the nature of latent cognitive processes (Doctoral thesis). University of California, Irvine. ProQuest ID: Baribault\_uci\_0030D\_15781. Merritt ID: ark:/13030/m5cc62b5. Retrieved from <https://escholarship.org/uc/item/31t8r1k3>
- Baribault, B., Donkin, C., Little, D., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Meta-studies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 2607–2612.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2), 137–163.
- DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating psychological science with metastudies: A demonstration using the risky-choice framing effect. *Perspectives on Psychological Science*, 17(6), 1704–1736. <https://doi.org/10.1177/17456916221079611>
- Kenett, R. S., & Rubinstein, A. (2021). Generalizing research findings for enhanced reproducibility: An approach based on verbal alternative representations. *Scientometrics*, 126, 4137–4151. <https://doi.org/10.1007/s11192-021-03914-1>
- Rubinchik, N. (2019). A demonstration of the meta-studies methodology using the risky-choice framing effect (Master’s thesis). Ohio State University. OhioLINK Electronic Theses and Dissertations Center. Retrieved from [http://rave.ohiolink.edu/etdc/view?acc\\_num=osu1574201911927335](http://rave.ohiolink.edu/etdc/view?acc_num=osu1574201911927335)
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128.
- Watts, D. (2017). Should social science be more solution-oriented? *Nature Human Behavior*, 1, 15. <https://doi.org/10.1038/s41562-016-0015>

## Dimensional versus conceptual incommensurability in the social and behavioral sciences

Eugene Vaynberg , Kate Nicole Hoffman , Jacqueline Mae Wallis  and Michael Weisberg\* 

Department of Philosophy, University of Pennsylvania, Philadelphia, PA, USA  
[eugenev@sas.upenn.edu](mailto:eugenev@sas.upenn.edu)  
[hoffmakn@sas.upenn.edu](mailto:hoffmakn@sas.upenn.edu)  
[jacqwa@sas.upenn.edu](mailto:jacqwa@sas.upenn.edu)  
[weisberg@phil.upenn.edu](mailto:weisberg@phil.upenn.edu)  
<https://philosophy.sas.upenn.edu/people/eugene-vaynberg>  
<https://philosophy.sas.upenn.edu/people/kate-nicole-hoffman>  
<https://philosophy.sas.upenn.edu/people/jacqueline-mae-wallis>  
<https://philosophy.sas.upenn.edu/people/michael-weisberg>

\*Corresponding author.

doi:10.1017/S0140525X23002182, e64

### Abstract

This commentary analyzes the extent to which the incommensurability problem can be resolved through the proposed alternative method of integrative experiment design. We suggest that, although one aspect of incommensurability is successfully addressed (*dimensional incommensurability*), the proposed design space method does not yet alleviate another major source of discontinuity, which we call *conceptual incommensurability*.

The concept of a design space is Almaatouq et al.'s major and important contribution to solving the incommensurability problem that arises in the social and behavioral sciences. The incommensurability problem, which the authors claim is caused in part by the promotion of "one-at-a-time" experiments that are conducted in theoretical isolation from other relevant experiments, has resulted in "irreconcilable theories and empirical results" (target article, sect. 1, para. 5). The kind of incommensurability at issue here is the inability to compare the same effect of interest across separate experiments. Call this *dimensional commensurability*. To address this, the design space's core features include (i) identification of plausibly relevant dimensions of the phenomenon of interest and (ii) assignment to each possible experiment a coordinate based on the dimensions the experiment is designed to investigate. Commensurability between experiments is thus supposed to be "baked in," since all experiments directed at answering a given question can be compared in the design space, allowing for more nuanced theories that take into account varying dimensions and contexts. Although we agree that widespread use of a design space could help address such dimensional incommensurability, namely, the many variables that make social and behavioral theories particularly complex, our worry is that this strategy does not yet alleviate another major source of incommensurability: Conceptual discontinuity between research projects.

Implicit in the process of assigning design space coordinates to experiments is the assumption that each dimension will track the same concept across experiments. If an experiment  $E_1$  investigates some dimension  $d$  and another experiment  $E_2$  also claims to investigate  $d$ , then commensurability requires not only that  $d$  is identified

as a variable or effect of interest in both cases but also that  $d$  is conceptually identical in both  $E_1$  and  $E_2$ . In other words, merely using equivalent terms to refer to the same purported dimension  $d$  does not yet achieve conceptual identity. In seeking to assign the results of  $E_1$  and  $E_2$  to the design space, what justifies their respective locations? Taking their variables of interest at face value will yield one set of coordinate assignments. But if the concepts that underpin these variables are not the same, then their subsequent relation in the design space may be inaccurate or misleading. As has been highlighted elsewhere in the literature on experimentation in the social and behavioral sciences (e.g., Scheel, Tiokhin, Isager, & Lakens, 2021), investigating human behavior requires well-defined concepts to ensure that observations and measurements accurately and adequately capture the phenomena of interest. This kind of incommensurability, call it *conceptual incommensurability*, affects both the validity of any given experiment as well as the ability to effectively compare two different experiments.

To use one of the target article's examples, possible dimensions of the design space for experiments on "group synergy" will include individual-level traits such as "average skill," "social perceptiveness," and "cognitive style," as well as group-level variables including "communications technology" and "incentive structure." Mapping how these variables interact through a shared design space would considerably improve understanding of group synergy. But what the design space does not address is the ambiguity within singular concepts that both guides the research project ("group synergy") and defines its relevant parameters ("average skill," "social perceptiveness," etc.). If it turned out that a set of experiments, each purporting to test the effect of social perceptiveness on group synergy, was differentially conceptualizing social perceptiveness, then it is not clear how the results of these experiments could be commensurable. To put it simply, can research group A be sure that research group B conceptualizes "social perceptiveness" in the same way? Plotting a set of experiments in a design space can obscure underlying conceptual discontinuities. There is reason to believe that research conducted under different guiding strategies might inherit conceptual discordance from the outset, precluding the construction of a design space in the first place (Lacey, 2005). Even prior to running experiments, the process of designing a space of possible experiments assumes that researchers will subsequently operate using those same concepts. This may turn out to be true, but not by fiat. Thus the design space, while addressing one aspect of the incommensurability problem (whether the *stated* effects of interest are the same) overlooks another (whether the *concepts* are equivalent).

Our concern is that the conceptual identity of the variables determining the design space must be ensured, not merely taken for granted. Without such conceptual identity, results are not guaranteed to be commensurable, and social and behavioral evidence will not be reconciled across experiments. Disagreements in two measurements cannot be resolved if the source of the discrepancy is unclear. Is it a problem in the measurement methodology? Is it a problem in the accuracy of the measurements themselves? Or is the problem that the measurements are either not capturing the relevant concept or not measuring the same concept across different studies? Feest (2022), for instance, argues that experimental psychologists must address three distinct reactivity challenges – all related to the ways psychological subject matter have dispositions to react to experimental contexts – in designing their experiments; Almaatouq et al.'s call for integrative experiment design across a design space suggests another challenge to designing experiments where genuine results can be distinguished from artifacts.

Thus, our suggestion is that, for commensurability to obtain in the design space as the authors advocate, researchers must first precisely specify the operative concepts. Such specification is challenging, as it involves what philosophers, historians, and scientists have called the problem of coordination: Coordinating measurement of directly observable entities with quantities of interest that can only be inferred from the observable entities (Kellen, Davis-Stober, Dunn, & Kalish, 2021). Since there is good reason to think that the specification of concepts used across the social and behavioral sciences does not occur on a systematic basis (e.g., Bringmann, Elmer, & Eronen, 2022; Scheel et al., 2021), the use of concepts to generate the proposed design space will inherit the same conceptual incommensurability.

**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References


- Bringmann, L. F., Elmer, T., & Eronen, M. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, 31(4), 340–346. <https://doi.org/10.1177/09637214221096485>
- Feest, U. (2022). Data quality, experimental artifacts, and the reactivity of the psychological subject matter. *European Journal for Philosophy of Science*, 12(1), 13. <https://doi.org/10.1007/s13194-021-00443-9>
- Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, 16(4), 767–778. <https://doi.org/10.1177/1745691620974771>
- Lacey, H. (2005). *Values and objectivity in science: The current controversy about transgenic crops*. Lexington Books.
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16(4), 744–755. <https://doi.org/10.1177/1745691620966795>

## Author's Response

### Replies to commentaries on beyond playing 20 questions with nature

Abdullah Almaatouq<sup>a\*</sup> , Thomas L. Griffiths<sup>b</sup> ,

Jordan W. Suchow<sup>c</sup> , Mark E. Whiting<sup>d</sup> ,

James Evans<sup>e,f</sup>  and Duncan J. Watts<sup>g</sup> 

<sup>a</sup>Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA; <sup>b</sup>Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA; <sup>c</sup>School of Business, Stevens Institute of Technology, Hoboken, NJ, USA; <sup>d</sup>School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA; <sup>e</sup>Department of Sociology, University of Chicago, Chicago, IL, USA; <sup>f</sup>Santa Fe Institute, Santa Fe, NM, USA and <sup>g</sup>Department of Computer and Information Science, Annenberg School of Communication, and Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA, USA  
[amaatouq@mit.edu](mailto:amaatouq@mit.edu)  
[tomg@princeton.edu](mailto:tomg@princeton.edu)  
[jws@stevens.edu](mailto:jws@stevens.edu)  
[markew@seas.upenn.edu](mailto:markew@seas.upenn.edu)  
[jevans@uchicago.edu](mailto:jevens@uchicago.edu)  
[djwatts@seas.upenn.edu](mailto:djwatts@seas.upenn.edu)

\*Corresponding author.

doi:10.1017/S0140525X23002789, e65

## Abstract

Commentaries on the target article offer diverse perspectives on integrative experiment design. Our responses engage three themes: (1) Disputes of our characterization of the problem, (2) skepticism toward our proposed solution, and (3) endorsement of the solution, with accompanying discussions of its implementation in existing work and its potential for other domains. Collectively, the commentaries enhance our confidence in the promise and viability of integrative experiment design, while highlighting important considerations about how it is used.

## R1. Overview

We are grateful to our colleagues for the effort they devoted to writing so many interesting and thoughtful commentaries, and we appreciate the diverse viewpoints conveyed therein. Our commentators not only raised a number of criticisms and concerns, but also offered constructive suggestions for the theoretical foundation of the integrative experiment design framework and the challenges faced when implementing it in practice. Although not everyone agrees with our characterization of the problem or with the solution we describe, the overall response to the target article reinforces our confidence that the problem we have identified is important and that the integrative approach is promising and practical, while highlighting important considerations about its use.

Given the breadth and nuance of the commentaries, addressing every point raised is not possible. Therefore, we have instead organized our response around three central themes that arose repeatedly across the commentaries. The first theme addresses those that challenge our characterization of the problem – specifically, the premise that integrating diverse experimental findings often proves inefficient or fails altogether due to the incommensurability inherent in the “one-at-a-time” experiment design. The second theme comprises commentaries that more or less agree with our problem identification but are skeptical of the proposed solution, raising arguments against its theoretical possibility, operational practicality, or effectiveness in addressing the problem. Notably, some commentaries in this category also offer potential solutions to the issues raised. The final theme includes responses that accept the proposed solution and also describe projects that embody the integrative approach or examine its potential application in other domains and fields of study.

## R2. Is there even a problem?

We begin by examining the extent to which the commentaries agree with our premise that the experimental social and behavioral sciences are often not cumulative, and that existing mechanisms for integrating disparate experimental findings do not work. We attribute this failure to the problem of incommensurability, where experiments are conducted in theoretical isolation from other relevant experiments, exacerbated by the one-at-a-time approach.

While the vast majority of commentaries expressed substantive agreement with our claim and diagnosis of the root cause, a few did not. Among those that disagreed, commentators' positions ranged from complete rejection of the target article's central premise to the suggestion that our outlook is overly optimistic and that there is, in fact, no hope for a cumulative tradition.



### R2.1. What problem?

Two commentaries argue that the integration-related issues highlighted in the target article are either unproblematic or do not warrant a shift in practices.

**Kellen, Cox, Donkin, Dunn, & Shiffrin (Kellen et al.)** adopt an assertive stance, disputing our critique of the one-at-a-time approach on the basis that (1) it seems to be working in some domains (e.g., human memory), where theory and one-at-a-time experiments have accumulated to form a self-consistent and empirically validated body of knowledge; and (2) inconsistent results within a literature do not necessarily indicate a failure of the one-at-a-time approach, as they may have alternative explanations. For example, inconsistent results might arise from hidden preconditions or moderators, or from studying different phenomena not explainable by a common theory.

We acknowledge that an observed lack of integration across several domains of interest does not imply that successful integration cannot occur in principle, or even that it has not occurred in practice. We discussed this in the target article and offered examples of successful instances such as mechanism design applied to auctions. As noted in the target article, the one-at-a-time approach seems to work well in domains characterized by low causal density, but other factors could influence its effectiveness, such as experimenters' knowledge about the range of relevant parameter values, the plausible range of changes in the outcome, and the nature of questions being asked – whether technological ( $x$  can do  $y$ ) or substantively theoretical ( $x$  is the mechanism that generates phenomena of interest  $y$ ). For more discussion on this point, see the related discussion by Meehl (1990) on why significance testing has worked sufficiently in agronomy, but not psychology.

**Kellen et al.** point to working memory as “an exemplary case” of a successful cumulative tradition, and indeed it is among the clearest examples of a field being able to better link results across one-at-a-time studies through reliance on shared experimental paradigms (e.g., free recall, the Deese–Roediger–McDermott (DRM) task, and change detection). But even this high-paradigm field has struggled with problems of integration. Consider Oberauer et al. (2018), a paper cited by Kellen et al. as evidence of the field's wealth of empirical findings, which begins with a premise very much like our own: “Any mature field of research in psychology – such as short-term/working memory – is characterized by a wealth of empirical findings. It is currently unrealistic to expect a theory to explain them all; theorists must satisfice with explaining a subset of findings.” Oberauer et al.'s proposed solution was the curation of benchmark datasets, whereas the integrative approach might be reasonably thought of as the *collection* of benchmark datasets. A recent attempt to collect such benchmark datasets in the context of working memory by Huang (2023), which was explicitly inspired by the integrative approach we propose, has made considerable progress toward developing unified theories of working memory as a result (Suchow, 2023).

In response to the second point, the inability of the one-at-a-time approach to uncover preconditions and moderators and to delineate between theoretically (and empirically) distinct regions of the space is precisely the failure mode that Newell highlighted in his article, with which we introduced our target article. Therefore, what **Kellen et al.** see as alternative explanations to the failings of the one-at-a-time approach are what we see as downstream symptoms of the target article's central premises. A

primary goal of the proposed integrative approach is to address these specific issues.

**Baron** contends that the one-at-a-time paradigm is concerned with demonstrating the existence of effects and causal chains, not their generality. However, if one accepts the notion that many of the social and behavioral phenomena are causally dense (“crud factor” by Meehl or “no true zeros” by Gelman), then it follows that almost every plausible hypothesized effect (i.e.,  $X \rightarrow Y$  relationship) exists to at least some degree, under some conditions, some of the time. If this is the case, then focusing on a single hypothesis and demonstrating its existence to at least some degree, under some (often unarticulated) conditions, some of the time, demonstrates nothing beyond what we already had good reason to believe. This line of reasoning is often used to criticize null hypothesis testing for theory evaluation: Anything that plausibly could have an effect will not have an effect that is exactly zero because of the high-causal density of our subject matter (Gelman, 2011; Meehl, 1967). In contrast, our argument in the target article is that experimental design should address two key questions: (a) how do the many plausible hypothesized effects contribute to the outcome of interest, both individually and in combination with other plausible hypothesized effects; and (b) how does this contribution vary depending on relevant contextual variables? This line of inquiry forms the core of the integrative approach to experiment design.

### R2.2 The solution is already here!

**Holleman, Dhimi, Hooge, & Hessels (Holleman et al.)** contest the lack of a realistic alternative to the one-at-a-time approach that could facilitate the integration of experimental findings. They point to “representative design,” a method introduced by Egon Brunswik about 70 years ago – nearly two decades before Newell's critique of the one-at-a-time approach. Representative design involves generating a sample of stimuli, either directly extracted from the environment or designed to retain its characteristics, to be representative of the population of environments to which the experimenter wishes to generalize. The integrative approach is indeed inspired by and builds upon Brunswik's representative design, which we cite in the target article. We identify at least two connections with his work. First, a frequent criticism of Brunswik's representative design revolves around the challenge of defining the universe of potential environments and formally sampling situations from it. The integrative approach directly addresses these challenges, as we discuss in the definition of the design space and sampling strategies (sects. 3.1 and 3.2 in the target article; sect. R3.2). Second, Brunswik's philosophy emphasizes sampling conditions from the agent's “natural environment.” This ensures that the conditions chosen for experimentation are representative of those to which the agent has adapted and to which generalizations would be applicable (this corresponds with the target article's Fig. 2C; what we call the region of ecological validity in the design space). While this strategy is suitable for certain scientific goals, we describe in section R3.2 other equally legitimate scientific pursuits that would involve sampling conditions that are either infrequently encountered or do not currently exist in the real world. Finally, we note that while representative design indeed bears some resemblance to the integrative approach, the former has not been widely adopted in the ensuing 70 years; thus, the need for an operational solution remains unmet.

### R2.3. *There is no hope*

**Mandel** objects that many phenomena in social and behavioral sciences may exhibit such extreme causal density as to defeat any attempts to generalize; hence, the target article understates the severity of the problem. We discussed exactly this possibility, noting that when one “point” in the (latent) space fails to provide information about any other point (including, as **Mandel** posits, the same point over time; similar issues were raised by **Olsson & Galesic**), any kind of generalization is unwarranted due to extreme sensitivity with respect to contextual factors (including time). As we noted in the target article, such an outcome would be disappointing to many, as it would essentially vitiate the potential for generalizable theory in that domain; however, it would not invalidate the integrative approach. On the contrary, it demonstrates its potential to reveal fundamental limits to prediction and explanation (**Hofman, Sharma, & Watts, 2017**; **Martin, Hofman, Sharma, Anderson, & Watts, 2016**; **Watts et al., 2018**). If true, it is surely preferable to characterize such limits than to indulge in wishful thinking and social science fiction. Moreover, applied research might still have merit, potentially by centering on the exact point (time, context, population) of interest (**Manzi, 2012**). This could yield reproducible social technologies even if not broad-based scientific theories. As we also note, however, such unforgiving cases are not a foregone conclusion. Rather, the extent to which they arise is itself an empirical question that the integrative approach is positioned to address. By conducting a sufficient number of integrative experiments across various domains, the approach could potentially lead to a “meta-metatheory” that clarifies under which conditions we can or cannot expect to identify generalizable findings.

### R3. *Is the integrative approach viable?*

Most commentaries broadly agreed with the target article’s framing of the problem and focused their discussion on the viability of the integrative approach as a potential solution. The target article describes the integrative experiment design approach as involving three key steps: (1) Explicit definition of an  $n$ -dimensional design space representing the universe of relevant experiments for a phenomenon; (2) judiciously sampling from this design space in alignment with specific goals; and (3) integrating and synthesizing the results through the development of theories that must address the heterogeneity (or invariance) of outcomes across this space. In this section, we review the commentaries related to each of these steps and then address broader concerns about the potential impact of the integrative approach on who participates in science.

#### R3.1. *The feasibility of constructing a design space*

Several commentaries challenge the first step of the integrative experimentation approach, arguing that constructing the “design space” (or “research cartography”) is practically difficult, if not theoretically impossible. Yet, some commentaries also offer possible solutions and suggestions.

**Olsson & Galesic** pose the important question: Where do the dimensions of the design space come from? They argue that drawing from past studies might lead to a biased representation of the true design space, influenced by implicit or explicit theories of the original researchers, methodological constraints, or adherence to a particular experimental paradigm. **Clark, Isch, Connor, & Tetlock (Clark et al.)** add that researchers may overlook certain

dimensions that challenge their previous work, contradict their favored theories, fall outside their expertise, or stem from the list of “socially off-limits” – yet plausible – dimensions. **Primbs, Dudda, Andresen, Buchanan, Peetz, Silan, & Lakens (Primbs et al.)** note that excluding any factor from the design space could lead to the dismissal of integrative experiments’ conclusions due to a crucial missing moderator. Such discourse gives rise to the question **Shea & Woolley** pose – “who decides what variables are included or receive more attention?” – and shares **Tsvetkova’s** concerns that this approach could worsen existing inequalities and hierarchies within academia, which we discuss further in section R3.4.

Shifting focus from the source of the dimensions to the dimensions themselves, **Gollwitzer & Prager** argue that these dimensions are often theoretical constructs, and the various ways of conceptualizing and operationalizing them may potentially lead to very large (if not infinite) design spaces. **Vaynberg, Hoffman, Wallis, & Weisberg** underscore the problems posed by a lack of precise operative concepts in the social and behavioral sciences, which could lead different researchers to conceptualize the same dimension differently, thereby creating “conceptual incommensurability” across experiments that ostensibly deal with the same theoretical construct. **Higgins, Gillett, Deschrijver, & Ross** add yet another layer of complexity by pointing out that even with the same conceptualization of a dimension, researchers might employ different instruments for measurement, leading to “measurement incommensurability” across experiments. They also discuss the implications of the validity and reliability issues of these measurement tools on the integrative approach. Finally, **Dubova, Sloman, Andrew, Nassar, & Musslick**, as well as **Necip Tunç & Tunç**, voice concerns that committing to any predetermined list of dimensions could be prematurely restrictive, potentially stifling the exploration of new dimensions beyond that list.

In addition to raising concerns, our commentators also offer some solutions. **Clark et al.**, as well as **Amerio, Coucke, & Cleeremans**, propose adversarial collaborations, wherein teams would include scholars who have previously published from multiple competing theoretical perspectives, possibly incorporating academics who study similar phenomena from various, even numerous, formerly competing standpoints. Requiring researchers to collaborate with theoretical adversaries would increase the likelihood of the research design space incorporating relevant dimensions. Rather than a winner-takes-all competition between seemingly contradictory hypotheses, this model would encourage exploration for genuine metatheories that clarify contexts in which different claims hold, resolving apparent discrepancies between leading scholars’ favored theories. **Primbs et al.** propose consensus meetings, wherein researchers convene to discuss and commit a priori to the list of dimensions, their operationalization, and the validity and reliability of chosen instruments. Such a consensus would also cover the implications that the results of integrative experiments have for their hypotheses, making it more challenging to dismiss the conclusions drawn from these experiments. Complementing adversarial collaborations and consensus meetings, **Katiyar, Bonnefon, Mehr, & Singh** suggest a high-throughput natural description approach to tackle the “unknown unknowns” in a specific design space. This approach involves systematically collecting and annotating large, representative corpora of real-world stimuli, leveraging mass collaboration, automation, citizen science, and gamification.

We are sympathetic to the concerns raised and appreciative of the constructive suggestions. As we acknowledge in the target

article, building the design space is not a solved problem, nor is it a single problem with a singular solution. In practice, the specific methods are yet to be worked out in detail and are likely to vary from one application to another, but adversarial collaborations, consensus meetings, and high-throughput approaches all seem like promising candidates. Fortunately, as we pointed out in the target article, the integrative approach does not require a design space with fixed, predetermined dimensions. On the contrary, the design space's dimensionality should remain fluid. For example, it can expand when identical points in the design space produce systematically varying results, indicating a need to actively search for an additional dimension to account for these differences. Alternatively, the design space can contract when experiments with systematic variations within the design space yield similar outcomes, suggesting that the dimensions in which they differ are irrelevant to the phenomenon of interest and therefore should be collapsed or omitted. In this way, concerns about misspecification, while reasonable, are not problematic for the integrative approach on their own. As we note in the target article, "the only critical requirement for constructing the design space is to do it explicitly and systematically by identifying potentially relevant dimensions (either from the literature or from experience, including any known experiments that have already been performed) and by assigning coordinates to individual experiments along all identified dimensions." Beyond that, we are agnostic about the specifics and look forward to seeing how best practices evolve through experience.

We also note that the issues raised in most of these commentaries also apply to the traditional one-at-a-time approach. Employing an integrative approach does not change, for example, the imprecision inherent in the field nor does it make variation across a behavioral dimension any more infinite than it already is. Thus, while we agree that additional advances are needed and care should be taken in how we design and execute our experiments, we do not see the current state of affairs as a weakness of the integrative approach so much as a weakness of all empirical work.

### R3.2. The challenge of effective sampling

As mentioned in the target article, efficient and effective sampling of the design space is a practical necessity given the limited resources available for conducting experiments. **Vandekerckhove** highlights an often-overlooked advantage of sampling from design spaces: Statistical efficiency. Efficiency can be achieved by increasing the number of sampled experimental conditions while decreasing the number of trials per experiment, thereby maximizing estimation accuracy, when the effect is heterogeneous (DeKay, Rubinchik, Li, & De Boeck, 2022).

The target article outlined different strategies and emphasized that *the choice of the best strategy is goal-dependent*. **Holleman et al.** and **Tsvetkova** note that some sampling strategies could select context and population combinations – or "environments" in Brunswik's terminology – that are not found or are impossible in reality, hence lacking ecological validity. They propose sampling following Brunswikian principles of representative design. Indeed, we agree that if the goal of research is to generalize results to the situations where the participant population functions, we can ensure ecological validity by selecting experimental conditions representative of those situations. It is worth noting, however, that what is considered ecologically valid or relevant can vary across populations, places, and periods, and thus representation of the "natural environment" can change. For instance,

**Salganik, Dodds, and Watts' (2006)** experimental music market may be more representative of today's environment, with the ubiquity of online audio streaming services like Spotify and SoundCloud, than when the study was initially conducted. Thus, there is merit in sampling beyond what is considered representative of a specific population, place, time, or situation. Conditions change and can be changed; conditions that might not currently exist in the "natural environment" can still be the foundation for valuable technological and platform designs; for example, those that modify the "natural environment" of a digital social network platform to elicit desired behaviors from users.

**Gollwitzer & Prager** argue against random sampling from the design space, calling for strategies that acknowledge a hierarchy in the potential experiments' informativeness and discriminability. They propose prioritizing experiments that, either logically or theoretically, hold more importance over more peripheral ones. While incorporating such knowledge in the sampling strategy is compatible with the target article's proposed approach, we refer to **Dubova, Moskvichev, and Zollman's (2022)** and **Musslick et al.'s (2023)** work, which examined the epistemic success of theory-driven experimentation strategies such as verification, falsification, novelty, and crucial experimentation, finding that if the objective is to uncover the underlying truth, random sampling proves to be a very robust strategy. We also note that adversarial collaborations and consensus meetings, proposed above, offer potentially useful mechanisms for aligning the sampling strategy with the research goal.

### R3.3. Contemplating the nature of "theory"

In the target article, we posit that, much like the traditional one-at-a-time paradigm, the goal of integrative experiment design is to cultivate "a reliable, cohesive, and cumulative theoretical understanding."

Several commentaries contest this claim about the integrative approach, suggesting it may be blindly empirical or even antitheoretical. In particular, many commentaries expressed discomfort at the suggestion that fitting a machine-learning model (such as a surrogate model within the "active learning" process) can generate or contribute to theory. However, as indicated by **Devezer** and reflected in those commentaries, there is no consensus on what a theory is or how to differentiate a good theory from a bad one. For instance, **Gal, Sternthal, & Calder** insist that theory should be expressed in terms of theoretical constructs, not variables. **Devezer**, as well as **Hoffman, Quillien, & Burum (Hoffman et al.)** and **Smaldino** demand theories to be mechanistic, identifying the underlying causal processes. **Hoffman et al.** stipulate that these mechanisms should be "well-grounded," meaning they can be explained in terms of well-understood processes. **Hullman** call for better a definition of what makes a theory interpretable. **Smaldino** argues that individual theories should align with a broader collection of related theories and share a set of common core assumptions, while **Baron** advocates for theories with broad scope, reflecting diverse, often ostensibly unrelated phenomena. **Vandekerckhove, Smaldino, and Olsson & Galesic** call for more computational models and quantitative theories. **Vandekerckhove** adds that ideally, these would take the form of likelihood functions – functions that describe the probability of data patterns under a theory – over the method space, while **Kummerfeld & Andrews** argue for using causal-discovery artificial intelligence (AI) to generate multivariate structural causal models.



We empathize with these perspectives on the essential features of a theory. We acknowledge that theories can originate from various sources, including historical records, ethnographic observations, everyday anecdotal experiences, data mining, fitting a black box machine-learning model, constructing analytically tractable mathematical models, or simply thinking hard about why people behave as they do. Nevertheless, a theory's truth is not guaranteed by satisfying any or all of these features, or by the method of its generation. It is also not clear why any specific feature or process of theorizing should always take precedence over others. Indeed, when pressed, researchers who advocate these features for theories will likely assert a connection between the presence of their preferred features and the "success" of the theory in explaining existing observations and predicting unseen data. Hence, we view the following as the most inclusive criteria for what we expect from theories: They should (a) accurately explain observed experiment results and (b) make accurate predictions about unseen experiments. To be clear, we do not object to theories incorporating any other features. However, we argue that a theory's primary evaluation criterion should be its accuracy in predicting unseen experimental outcomes, including out-of-distribution and underintervention scenarios.

More broadly, the commentaries have caused us to reflect more deeply on our use of the word "understanding" when describing the ultimate goal of integrative experimentation. A better word would have been "explanation," which we view as having two subgoals: (1) Providing an accurate representation of how the world works (emphasizing causality, generalizability, prediction, etc.), and (2) resonating with a human interpreter by evoking a sensation of understanding (emphasizing interpretability and sense making). We sense from some of the commentaries that these two goals are often conflated: *Accurate theories are also the ones that make sense to a human interpreter*. However, in our view these objectives are distinct in theory and often conflict in practice. This conflict stems from the high-causal density of phenomena under study, leading us to a situation where accurate theories (with robust out-of-distribution performance) may not satisfy our intuitive understanding due to their complexity – the number of factors and scale of their interactions. Conversely, understandable explanations are often incorrect, as simple theories cannot capture the complexity of the phenomenon. Perhaps what our commentators find objectionable about our approach is its preferencing of accuracy over sense making. In other words, the integrative approach is about theory, but it presents a version of theory that places second what many consider theory's primary purpose: Generating a subjective sense of understanding.

Looking ahead, we might have to embrace an era of machine-generated or machine-aided social scientific theories, especially if we prioritize accuracy over sense making in domains of high-causal density, a line of thinking we will explore further in future publications.

#### R3.4. Diversity and power dynamics

Shea & Woolley, as well as Tsvetkova, raised concerns that the integrative approach might centralize power among a small number of elite researchers and well-funded institutions, exacerbating existing inequalities in research. Shea & Woolley criticize our use of high-energy physics as an exemplar of successful large-group collaborations, noting that that fields requiring significant infrastructure investment tend to be more hierarchical and face

significant issues such as sexual harassment and gender-participation gaps.

In response to these concerns, we provide two clarifications. First, while we agree that diversity, equity, and inclusivity must be first-order considerations in any scientific reform, we see them as distinct issues from the intellectual merits of the argument for an integrative approach. Second, there are several mechanisms to ensure that a diversity of perspectives is heard. In the target article (see sects. 5.7 and 5.9) we highlighted the ways in which the integrative approach could allow a broader range of contributors to be involved in the process of designing, conducting, and analyzing experiments. The suggestions raised in the commentaries – such as adversarial collaborations and consensus meetings (see sect. R3.1) – provide further opportunities for participation.

In general, the responsibility of facilitating wide-ranging contributions to science is a burden shared by the entire scientific community. The integrative approach offers a distinctive set of opportunities and challenges in our collective pursuit of this goal: It not only expands the ways in which researchers can contribute to and benefit from research, but also potentially introduces new social dynamics we have to navigate to ensure this potential is achieved.

#### R4. Does the integrative approach have reach?

This last theme considers commentaries that endorse our proposed solution and illustrate projects that embody the integrative approach or explore its potential extensions to other areas and domains.

Cyrus-Lai, Tierney, & Uhlmann describe a recent crowd-sourced initiative that brings together several designs, analyses, theories, and data collection teams. This project further demonstrates the limitations of the one-at-a-time approach and champions the need to evaluate "many theories in many ways." Tsvetkova suggests that the integrative approach's initial step – "research cartography" – can consolidate knowledge and invigorate new research, retrospectively, beyond the prospective goals of integrative experiments. She envisions a Wikidata-style database that would contain all social and behavioral knowledge from experiments, enabling the identification of research gaps, established findings, and contentious issues. Li & Hartshorne highlight the theoretical advancements nestled between the traditional one-at-a-time approach and the "ideal" version of the integrative approach. They highlight the potential for studies that employ large and diverse sets of stimuli, encompass a broad demographic range of subjects, or engage a variety of related tasks – even without systematic exploration. Meanwhile, Simonton highlights the value of infusing the integrative approach with correlational methods.

Glaser, along with Ghai & Banerjee, make a strong case for expanding the integrative approach to within-subject designs. They stress its statistical efficiency and the significance of individual differences. Additionally, Haartsen, Gui, & Jones (Haartsen et al.) propose a method that combines Bayesian optimization with within-subject designs to further increase the efficiency of data collection.

Lastly, Haartsen et al. highlight the potential value of the integrative approach in domains such as psychiatry and cognitive development. Titone, Hernández-Rivera, Iniesta, Beatty-Martínez, & Gullifer extend this approach to evaluate the implications of bilingualism on the mind and brain. They point to ongoing work consistent with the integrative approach, drawing connections between the *systems framework of bilingualism* and research cartography. Dohrn & Mezzadri extrapolate the integrative approach to thought

experiments. While these commentaries are clear and persuasive, our lack of expertise in these domains prevents us from contributing further substance to these discussions. The message from these commentaries is that issues stemming from the incommensurability characteristic of the “one-at-a-time” experiment design extend beyond the social and behavioral sciences and that the integrative approach may be fruitful in these domains and others beyond them.

Overall, these discussions and proposals sketch a vibrant picture of the various ways the integrative approach can be applied and expanded across different contexts. They reinforce the value and potential of this methodology, offering much to consider and incorporate into our own research.

In closing, we express our deep appreciation to all the commentators for their thoughtful insights and stimulating discussion. We are eager to continue engaging with the research community, incorporating these valuable suggestions into our work, and collectively advancing the field of social and behavioral science research.

## References

- DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating psychological science with metastudies: A demonstration using the risky-choice framing effect. *Perspectives on Psychological Science*, 17(6), 1704–1736. <https://doi.org/10.1177/17456916221079611>
- Dubova, M., Moskvichev, A., & Zollman, K. (2022). Against theory-motivated experimentation in science. BITSS. <https://doi.org/10.31222/osf.io/ysv2u>
- Gelman, A. (2011). Causality and statistical learning. *The American Journal of Sociology*, 117(3), 955–966.
- Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324), 486–488.
- Huang, T. (2023). A quasi-comprehensive exploration of the mechanisms of spatial working memory. *Nature Human Behaviour*, 7(5), 729–739.
- Manzi, J. (2012). *Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society*. Basic Books.
- Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International conference on world wide web* (pp. 683–694). International World Wide Web Conferences Steering Committee. ISBN: 978-1-4503-4143-1.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195–244.
- Musslick, S., Hewson, J. T., Andrew, B. W., Strittmatter, Y., Williams, C. C., Dang, G., ... Holland, J. G. (2023). An evaluation of experimental sampling strategies for autonomous empirical research in cognitive science. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. Retrieved from <https://escholarship.org/uc/item/5ch569fg>
- Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., ... Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, 144(9), 885–958.
- Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854–856.
- Suchow, J. W. (2023). Scaling up behavioural studies of visual memory. *Nature Human Behaviour*, 7(5), 672–673.
- Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., Grubestic, A., ... Salganik, M. (2018). *Explanation, prediction, and causality: Three sides of the same coin?* <https://doi.org/10.31219/osf.io/u6vz5>