Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences

Authors

Abdullah Almaatouq[1], Thomas L. Griffiths[2], Jordan W. Suchow[3], Mark E. Whiting[4], James Evans[5], and Duncan J. Watts[4]

Affiliations

[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139

[2] Departments of Psychology and Computer Science, Princeton University, Princeton, NJ 08540

[3] School of Business, Stevens Institute of Technology, Hoboken, NJ 07030

[4] University of Pennsylvania, Philadelphia, PA 19104

[5] Department of Sociology, University of Chicago, Chicago, IL 60637; Santa Fe Institute, Santa Fe, NM 87501

*Correspondence to: amaatouq@mit.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. The 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.

Keywords: (in)commensurability, cumulative knowledge, generalizability, experiments

1. Introduction **3**

2. The "One-at-a-Time" Paradigm **6**
2.1 The problem with the one-at-a-time paradigm 7
2.2 The universe of possible experiments 8
2.3 Incommensurability leads to irreconcilability 12

3. From One-at-a-Time to Integrative by Design **13**
3.1 Constructing the design space 14
3.2 Sampling from the design space 17
3.3 Building and testing theories 20

4. Existing Steps towards Integrative Experiments **23**
4.1. Factors influencing moral judgments 24
4.2. The space of risky decisions 25
4.3. A metastudy of subliminal priming effects 26

5. Critiques and Concerns **28**
5.1. Isn't the critique of the one-at-a-time approach unfair? 28
5.2. Can't we solve the problem with meta-analysis? 30
5.3. How does it differ from other recent innovations in psychology? 31
5.4. What about unknown unknowns? 32
5.5. This sounds great in principle but it is impossible to do in practice 33
5.6. Even if such experiments are possible, costs will be prohibitive 36
5.7. Does this mean that small labs can't participate? 37
5.8. Shouldn't the replication crisis be resolved first? 38
5.9 This proposal is incompatible with incentives in the social sciences 39

6. Conclusion **41**

References **42**

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 twenty 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; J. R. 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 (N. J. Allen & Hecht, 2004; Hackman et al., 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 et al., 2000; Barron, 2003; Devine et al., 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 et al., 2002), share information about high-quality solutions (Mason & Watts, 2012), or correct errors (Mao et al., 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 test the effects of some independent variables on other dependent variables while suppressing innumerable "aspects of the situation."[1] Setting aside the much-discussed problems of replicability and reproducibility,

[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.

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; D. J. 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.[2]

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”

[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., 2017).

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.[3] 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

[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 et al., 2011) address the problem. While these proposals are worthy, their focus is on individual results, not on how collections of results fit together.

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 & Gachter, 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 et al., 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, 2020; 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 is 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 (D. J. 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; M. A. Turner & Smaldino, 2022; Yarkoni, 2020). Researchers working independently in the same domain of inquiry will therefore invariably make design choices (e.g., parameter settings, subject pools) differently (Breznau 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 put 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, et al., 2021; J. R. 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 U.S. 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.[4] (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.)

[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.

Figure 1 offers 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 et al., 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 et al., 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 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 et al., 2017; Yarkoni, 2020). 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, 2021).

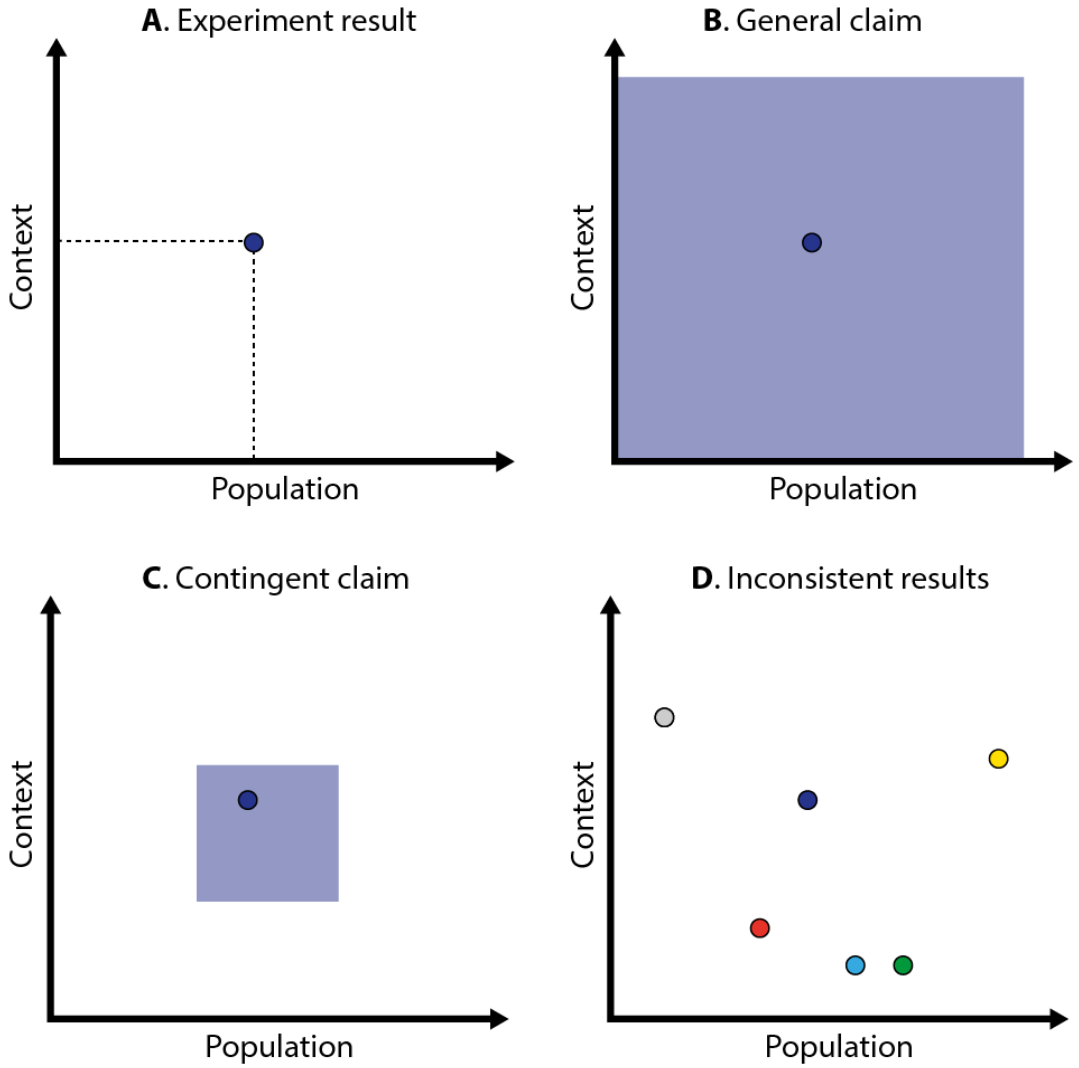

Figure 1. Implicit Design Space. Panel **A** depicts a single experiment (a single point) that generates a result in a particular sample population and context; the point’s color represents a relationship between variables. Panel **B** depicts the expectation that results will generalize over broader 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.

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 et al., 2016; D. J. Watts, 2017; Yarkoni, 2020).

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 et al., 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 et al., 2022), and their heterogeneity (Szaszi 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 pre-analysis 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 pre-registered 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 fifty 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 (S. Larson & Martone, 2009; Rubin et al., 2006; J. A. 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 et al., 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 et al., 2015), social perceptiveness (Engel et al., 2014; Kim et al., 2017; Woolley et al., 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 et al., 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 upwards of fifty 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.[5]

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

[5] To illustrate with another example, cultural psychologists such as Hofstede (2001), 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 et al. (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.

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 et al., 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 become 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 et al., 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.[6] Active learning has become an important tool in the design of A/B tests in industry (Letham et al., 2019) and, more recently, of behavioral experiments in the lab (Balietti et al., 2020).[7] 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

[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 et al., 2012), materials and mechanical designs (Burger et al., 2020; Gongora et al., 2020; Lei et al., 2021), and chemical reaction screening (Eyke et al., 2020, 2021; Shields et al., 2021)—just to mention a few.

and sampling strategy is flexible, and the best alternative to choose will depend on the problem (Eyke et al., 2021).[8]

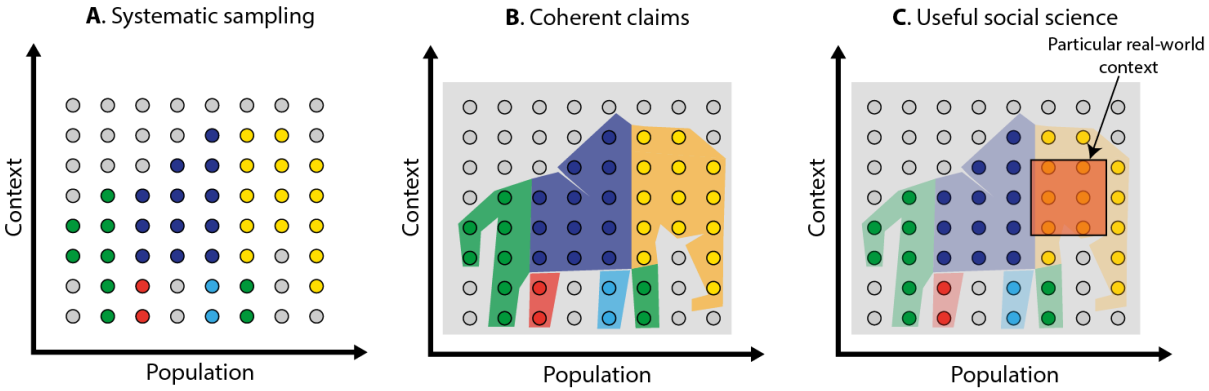

Figure 2. Explicit Design Space. Panel A shows that systematically sampling the space of possible experiments can reveal contingencies, thereby increasing the integrativeness of theories (as shown in Panel B). Panel C depicts that what matters most is the overlap between the most practically useful conditions 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.

We will not explore the details of these methods or their implementation,[9] as this large topic has been—and continues to be—extensively developed in the machine learning and statistics communities.[10] 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 Figure 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 Figure 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.

[8] For example, surrogate models can be probabilistic models (e.g., a Gaussian process) as well as non-probabilistic (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 et al., 2017).
[10] See Settles (2011), Greenhill (2020), and Ren et al. (2021) for surveys on active learning.

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 (D. Watts, 2017; D. J. Watts, 2014), 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, whereas 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; D. J. Watts, 2017). We now describe each of these adaptations in more detail.

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 and manifesting great variance across datasets (Geman et al., 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.[11] 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, 2020). 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 et al., 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, (c) *integrate* the newly

[11]Given that the data from the integrative approach is 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.

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 et al., 2017; Martin et al., 2016; D. J. Watts et al., 2018).[12] 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.

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

[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.

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 et al., 2022; Grubbs, 2022; Wiernik et al., 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; D. J. Watts, 2017).

4. Existing Steps towards 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 (Figure 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 longstanding 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 et al., 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 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 et al., 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 et al., 2019; Peterson et al., 2021).

By sampling the space of possible experiments (in this case, gambles) much more densely (Figure 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 et

al., 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 effect 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 and colleagues (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 (Figure 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 et al. (2022).

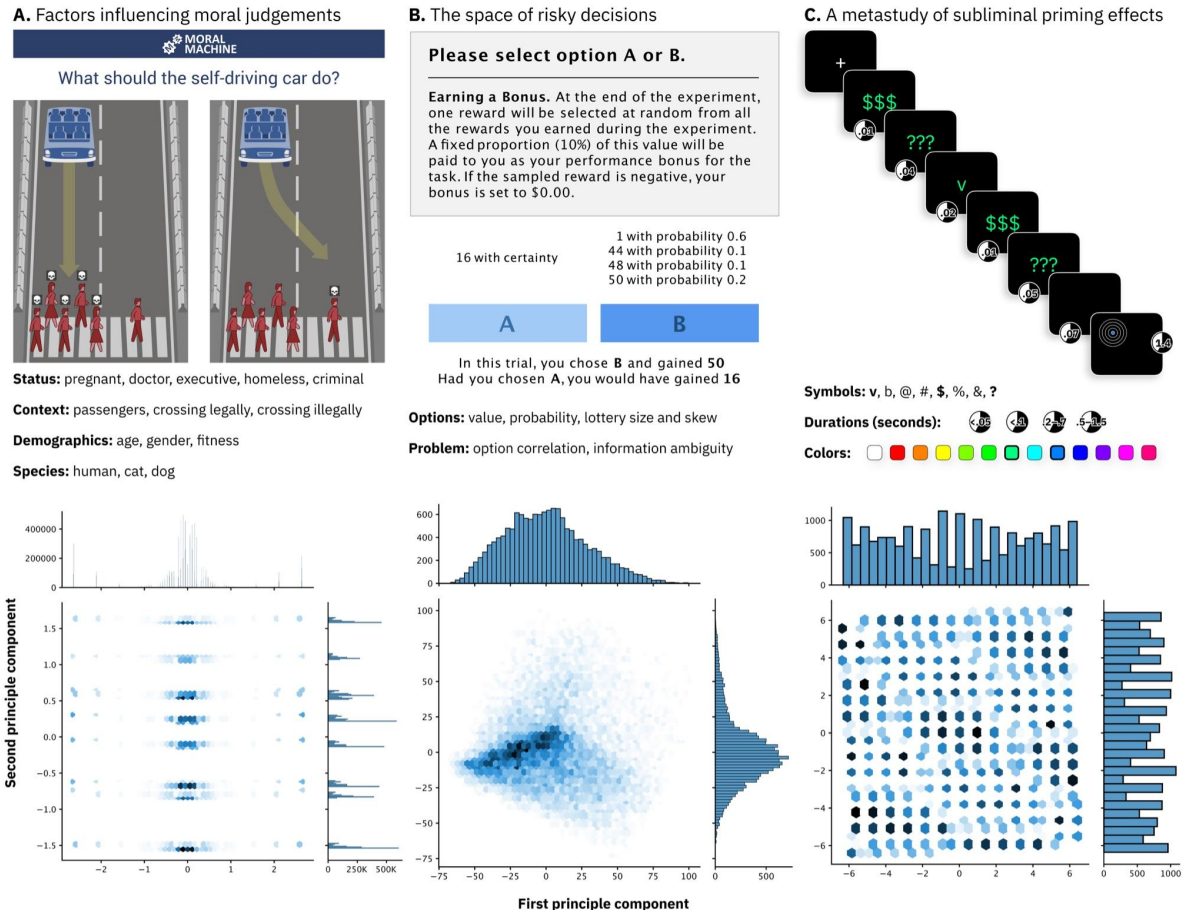

Figure 3. Examples of integrative experiments. The top row illustrates the experimental tasks used in the Moral Machine, decisions under risk, and subliminal priming effects experiments, respectively, followed by the parameters varied across each experiment (bottom row). Each experiment instance (i.e., 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.

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 science. 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 path breaking 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 is 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 Section 5.6 for a discussion about costs).

Second, whereas 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 et al., 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 et al., 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.).[13] 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

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

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 Section 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 hand, 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.

Virtual lab environments. Software packages such as jsPsych (de Leeuw, 2015) nodeGame (Balietti, 2017), Dallinger (https://dallinger.readthedocs.io/), Pushkin (Hartshorne et al., 2019), Hemlock (Bowen, n.d.), and Empirica (Almaatouq, Becker, et al., 2021) 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).

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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 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).

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.

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 non-experimental data.[14] In contrast, the current scale of experiments in the experimental social and behavioral sciences do 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 data sets, as in Agrawal et

[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 data sets (e.g., from social media and e-commerce platforms) are increasingly prevalent in political science, sociology, and economics.

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 multi-decade 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.[15] 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.[16] 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 meta-experiment may seem to exclude small labs as well as 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 meta-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.

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

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

non-replication 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, 2020). 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), etc.—have been obtained via collaborations of thousands of researchers.[17] 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 (L. Allen et al., 2014), a high-level taxonomy with 14 roles that describe typical contributions to scholarly output; the taxonomy has been adopted as an ANSI/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

[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.

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 non-standardized 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 pre-registered 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.

7. 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.

8. Funding statement. This work was supported in part by the Alfred P. Sloan Foundation (2020-13924) and the NOMIS Foundation.

9. Conflicts of Interest statement. None.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., Abdinov, O., Aben, R., Abi, B., Abolins, M., AbouZeid, O. S., Abramowicz, H., Abreu, H., Acharya, B. S., Adamczyk, L., Adams, D. L., Addy, T. N., Adelman, J., Adomeit, S., … 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., Adams, C., Adams, T., Addesso, P., Adhikari, R. X., Adya, V. B., Affeldt, C., Agathos, M., Agatsuma, K., Aggarwal, N., Aguiar, O. D., Aiello, L., Ain, A., Ajith, P., … 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*. 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. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(36). https://doi.org/10.1073/pnas.2101062118

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021).

Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*. 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.

Arrow, H., McGrath, J. E., & Berdahl, J. L. (2000). *Small Groups as Complex Systems: Formation, Coordination, Development, and Adaptation*. SAGE Publications.

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 Publications.

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., Bonnefon, J.-F., & 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., Daulton, S., Letham, B., Wilson, A. G., & Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*, *33*. https://research.fb.com/wp-content/uploads/2020/12/BOTORCH-A-Framework-for-Efficient-Monte-Carlo-Bayesian-Optimization.pdf

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. (2020). Optimal design of experiments to identify latent behavioral types. *Experimental Economics*. https://doi.org/10.1007/s10683-020-09680-w

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & 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., Bollen, K. A., Brembs, B., Brown, L., & Camerer, C. (2017). Redefine statistical significance. *Nature Human Behaviour*, 1.

Berkman, E. T., & Wilson, S. M. (2021). So Useful as a Good Theory? The Practicality Crisis in (Social) Psychological Theory. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 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.gitlab.io/hemlock

Brewin, C. R. (2022). Impact on the legal system of the generalizability crisis in psychology [Review of *Impact on the legal system of the generalizability crisis in psychology*]. *The Behavioral and Brain Sciences*, *45*, e7.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, 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; with results in physical and social perception*. *60*. https://psycnet.apa.org/fulltext/1949-00441-000.pdf

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., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S., & 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., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. In *Canadian Psychology/Psychologie canadienne* (Vol. 61, Issue 4, pp. 349–363). https://doi.org/10.1037/cap0000216

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … 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: A Journal of the Association for Psychological Science*, *9*(1), 40–48.

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. (2019). *The handbook of research synthesis and meta-analysis*. https://books.google.com/books?hl=en&lr=&id=tfeXDwAAQBAJ&oi=fnd&pg=PR5&dq=Vevea+J+L+Coburn+K+(2019)+Publication+bias+In+H+Cooper+L+V+Hedges+J+C+Valentine+(Eds+)+The+hand-+book+of+research+synthesis+and+meta-analysis+Russell+Sage+Foundation&ots=RMrBjsbI6W&sig=4StuMLVay62R04lJg001y4KJ4NM

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Debrouwere, S. (2020). *The conceptual, cunning and conclusive experiment in psychology*. https://users.ugent.be/~stdbrouw/2020-02-19-stijn-debrouwere-conceptual-cunning-and-conclusive-experiment.pdf

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: A Journal of the Association for Psychological Science*, 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: An Official Law Review of the University of Arizona College of Law and the University of Miami School of Law*, *7*(3), 622–727.

Devine, D. J., & Philips, 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 International Higher Education.

Dubova, M., Moskvichev, A., & Zollman, K. (2022). Against theory-motivated experimentation in science. In *BITSS*. https://doi.org/10.31222/osf.io/ysv2u

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … 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., & Gachter, 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*. 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., Alford, L. K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, M., Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … 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: The Official Journal of the Mathematical Association of America*, *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., Morgan, E. F., & 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.

Group, N. C. W., & Others. (2022). ANSI/NISO Z39. 104-2022, CRediT, Contributor Roles Taxonomy.[S. l.]. *Baltimore, MD: National Information Standards Organization*. https://www.niso.org/publications/z39104-2022-credit

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.

Hackman, J. R., Richard Hackman, J., & Morris, C. G. (1975). Group Tasks, Group Interaction Process, and Group Performance Effectiveness: A Review and Proposed Integration. In *Advances in Experimental Social Psychology* (pp. 45–99). https://doi.org/10.1016/s0065-2601(08)60248-8

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. In *Behavior Research Methods* (Vol. 51, Issue 4, pp. 1782–1803). https://doi.org/10.3758/s13428-018-1155-z

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*, *355*(6324), 486–488.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

Hofstede, G. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. https://digitalcommons.usu.edu/unf_research/53/

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*. *277*. 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., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, 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., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … 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*, 2316–2329.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … 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., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … 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. In *arXiv [stat.ML]*. arXiv. 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., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … 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., & Martone, M. (2009). Ontologies for neuroscience: What are they and what are they good for? *Frontiers in Neuroscience*, *3*. 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. In *Bayesian Analysis* (Vol. 14, Issue 2, pp. 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*.

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.

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.

Li, W., Germine, 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*. https://doi.org/10.1002/icd.2348

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.

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., Duguid, S. J., Egelkamp, C. L., Fichtel, C., Fischer, J., Flessert, M., Hanus, D., Haun, D. B. M., Haux, L. M., Hernandez-Aguilar, R. A., Herrmann, E., Hopper, L. M., Joly, M., Kano, F., … Watzek, J. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PloS One*, *14*(10), e0223675.

Manzi, J. (2012). Uncontrolled: The Surprising Payoff of Trial-and-Error for Business. *Politics, and Society. Basic Books*, 1–320.

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* (Proceedings of the 25th International Conference on World Wide Web No. 978-1-4503-4143-1; pp. 683–694). 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*, *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, Englewood Cliffs, NJ.

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., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., Brody, I., Chabris, C. F., Chang, E., Chapman, G. B., Dannals, J. E., Goldstein, N. J., Goren, A., Hershfield, H., Hirsch, A., … 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., Kay, J. S., Lee, T. W., Akinola, M., Beshears, J., Bogard, J. E., Buttenheim, A., Chabris, C. F., Chapman, G. B., Choi, J. J., Dai, H., Fox, C. R., Goren, A., Hilchey, M. D., … 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*, *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., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … 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., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

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. (2019). A problem in theory. *Nature Human Behaviour*. 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 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*, *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*, *372*(6547), 1209–1214.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., Cavanagh, J. F., & Erev, I. (2019). Predicting human decisions with behavioral theories and machine learning. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1904.06866

Preckel, F., & Brunner, M. (2017). *Nomological nets*. https://scholar.google.com/scholar?cluster=6872766253707406695&hl=en&oi=scholarr

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2021). A Survey of Deep Active Learning. *ACM Comput. Surv.*, *54*(9), 1–40.

Reuss, H., Kiesel, A., & Kunde, W. (2015). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62.

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., Sim, I., Chute, C. G., Storey, M.-A., Smith, B., Day-Richter, J., Noy, N. F., & Musen, M. A. (2006). National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. In *OMICS: A Journal of Integrative*

Biology (Vol. 10, Issue 2, pp. 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., Janey, J. M., Adams, R. P., & 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. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1206.2944

Steiner, I. D. (1972). *Group process and productivity*. Academic press New York.

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*.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & 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. National Acad Sciences.

Tasca, G. A. (2021). Team cognition and reflective functioning: A review and search for synergy. *Group Dynamics: Theory, Research, and Practice: The Official Journal of*

Division 49, Group Psychology and Group Psychotherapy of the American Psychological Association, *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., McCarthy, R. J., Riegelman, A., Silberzahn, R., & 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., Mernyk, J. S., Redekopp, C., Cashman, M., Submitters, Q. S. D., Druckman, J. N., Rand, D. G., & Willer, R. (2022). Megastudy identifying successful interventions to strengthen Americans' democratic attitudes. *Preprint*. https://doi.org/10.1098/rsos.181308

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*, 0015.

Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., Grubesic, A., Hofman, J. M., Rohrer, J. M., & 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*, *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. (2020). The Generalizability Crisis. *Behavioral and Brain Sciences*, 1–37.

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/vncwe

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.