Replies to commentaries on Beyond Playing 20 Questions with Nature

Authors

Abdullah Almaatouq[1], Thomas L. Griffiths[2], Jordan W. Suchow[3], Mark E. Whiting[4], James Evans[5], and Duncan J. Watts[6]

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] School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104

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

[6] Department of Computer and Information Science, Annenberg School of Communication, and Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA 19104

*Correspondence to: amaatouq@mit.edu

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

Overview

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

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

R1. Is there even a problem?

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

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

R1.1. What problem?

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

Kellen, Cox, Donkin, Dunn, and Shiffrin (Kellen et al.) adopt an assertive stance, disputing our critique of the one-at-a-time approach on the basis that (1) it seems to be working in some domains (e.g., human memory), where theory and one-at-a-time

experiments have accumulated to form a self-consistent and empirically validated body of knowledge; and (2) inconsistent results within a literature do not necessarily indicate a failure of the one-at-a-time approach, as they may have alternative explanations. For example, inconsistent results might arise from hidden preconditions or moderators, or from studying different phenomena not explainable by a common theory.

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

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

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

Baron contends that the one-at-a-time paradigm is concerned with demonstrating the existence of effects and causal chains, not their generality. However, if one accepts the notion that many of the social and behavioral phenomena are causally dense ("crud factor" by Meehl or "no true zeros" by Gelman), then it follows that almost every plausible hypothesized effect (i.e., $X \rightarrow Y$ relationship) exists to at least some degree, under some conditions, some of the time. If this is the case, then focusing on a single hypothesis and demonstrating its existence to at least some degree, under some (often unarticulated)

conditions, some of the time, demonstrates nothing beyond what we already had good reason to believe. This line of reasoning is often used to criticize null hypothesis testing for theory evaluation: anything that plausibly could have an effect will not have an effect that is exactly zero because of the high causal density of our subject matter (Gelman, 2011; Meehl, 1967). In contrast, our argument in the target article is that experimental design should address two key questions: (a) how do the many plausible hypothesized effects contribute to the outcome of interest, both individually and in combination with other plausible hypothesized effects; and (b) how does this contribution vary depending on relevant contextual variables? This line of inquiry forms the core of the integrative approach to experiment design.

R.1.2 The solution is already here!

Holleman, Dhami, Hooge, and Hessels contest the lack of a realistic alternative to the one-at-a-time approach that could facilitate the integration of experimental findings. They point to “representative design,” a method introduced by Egon Brunswik about 70 years ago—nearly two decades before Newell’s critique of the one-at-a-time approach. Representative design involves generating a sample of stimuli, either directly extracted from the environment or designed to retain its characteristics, to be representative of the population of environments to which the experimenter wishes to generalize. The integrative approach is indeed inspired by and builds upon Brunswik's representative design, which we cite in the target article. We identify at least two connections with his work. First, a frequent criticism of Brunswik’s representative design revolves around the challenge of defining the universe of potential environments and formally sampling situations from it. The integrative approach directly addresses these challenges, as we discuss in the definition of the design space and sampling strategies (Sections 3.1 and 3.2 in the target article; Section R.2.1 in the author response). Second, Brunswik’s philosophy emphasizes sampling conditions from the agent’s “natural environment.” This ensures that the conditions chosen for experimentation are representative of those to which the agent has adapted and to which generalizations would be applicable (this corresponds with the target article’s Figure 2C; what we call the region of ecological validity in the design space). While this strategy is suitable for certain scientific goals, we describe in R.2.2 other equally legitimate scientific pursuits that would involve sampling conditions that are either infrequently encountered or do not currently exist in the real world. Finally, we note that while representative design indeed bears some resemblance to the integrative approach, the former has not been widely adopted in the ensuing 70 years; thus, the need for an operational solution remains unmet.

R.1.3. There is no hope.

Mandel objects that many phenomena in social and behavioral science may exhibit such extreme causal density as to defeat any attempts to generalize; hence, the target article understates the severity of the problem. We discussed exactly this possibility, noting that when one “point” in the (latent) space fails to provide information about any other point (including, as Mandel posits, the same point over time; similar issues were raised by **Olsson and Galesic**), any kind of generalization is unwarranted due to extreme sensitivity with respect to contextual factors (including time). As we noted in the target article, such an outcome would be disappointing to many, as it would essentially vitiate the potential for generalizable theory in that domain; however, it would not invalidate the integrative

approach. On the contrary, it demonstrates its potential to reveal fundamental limits to prediction and explanation (Hofman et al., 2017; Martin et al., 2016; Watts et al., 2018). If true, it is surely preferable to characterize such limits than to indulge in wishful thinking and social science fiction. Moreover, applied research might still have merit, potentially by centering on the exact point (time, context, population) of interest (Manzi, 2012). This could yield reproducible social technologies even if not broad-based scientific theories. As we also note, however, such unforgiving cases are not a foregone conclusion. Rather, the extent to which they arise is itself an empirical question that the integrative approach is positioned to address. By conducting a sufficient number of integrative experiments across various domains, the approach could potentially lead to a "meta-metatheory" that clarifies under which conditions we can or cannot expect to identify generalizable findings.

R2. Is the integrative approach viable?

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

R.2.1. The feasibility of constructing a design space

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

Olsson and Galesic pose the important question: where do the dimensions of the design space come from? They argue that drawing from past studies might lead to a biased representation of the true design space, influenced by implicit or explicit theories of the original researchers, methodological constraints, or adherence to a particular experimental paradigm. **Clark, Isch, Connor, Tetlock (Clark et al.)** add that researchers may overlook certain dimensions that challenge their previous work, contradict their favored theories, fall outside their expertise, or stem from the list of "socially off-limits"—yet plausible—dimensions. **Primbs, Dudda, Andresen, Buchanan, Peetz, Silan, and Lakens (Primbs et al.)** note that excluding any factor from the design space could lead to the dismissal of integrative experiments' conclusions due to a crucial missing moderator. Such discourse gives rise to the question **Shea and Woolley** pose—"who decides what variables are included or receive more attention?"—and shares **Tsvetkova's** concerns that this approach could worsen existing inequalities and hierarchies within academia, which we discuss further in R.2.4.

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

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

We are sympathetic to the concerns raised and appreciative of the constructive suggestions. As we acknowledge in the target article, building the design space is not a solved problem, nor is it a single problem with a singular solution. In practice, the specific methods are yet to be worked out in detail and are likely to vary from one application to another, but adversarial collaborations, consensus meetings, and high-throughput approaches all seem like promising candidates. Fortunately, as we pointed out in the target article, the integrative approach does not require a design space with fixed, predetermined dimensions. On the contrary, the design space's dimensionality should remain fluid. For example, it can expand when identical points in the design space produce systematically varying results, indicating a need to actively search for an additional dimension to account for these differences.

Alternatively, the design space can contract when experiments with systematic variations within the design space yield similar outcomes, suggesting that the dimensions in which they differ are irrelevant to the phenomenon of interest and therefore should be collapsed or omitted. In this way, concerns about misspecification, while reasonable, are not problematic for the integrative approach on their own. As we note in the target article, “the only critical requirement for constructing the design space is to do it explicitly and systematically by identifying potentially relevant dimensions (either from the literature or from experience, including any known experiments that have already been performed) and by assigning coordinates to individual experiments along all identified dimensions.” Beyond that, we are agnostic about the specifics and look forward to seeing how best practices evolve through experience.

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

R.2.2. The challenge of effective sampling

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

The target article outlined different strategies and emphasized that *the choice of the best strategy is goal-dependent*. **Holleman et al.** and **Tsvetkova** note that some sampling strategies could select context and population combinations—or “environments” in Brunswik’s terminology—that are not found or are impossible in reality, hence lacking ecological validity. They propose sampling following Brunswikian principles of representative design. Indeed, we agree that if the goal of research is to generalize results to the situations where the participant population functions, we can ensure ecological validity by selecting experimental conditions representative of those situations. It is worth noting, however, that what is considered ecologically valid or relevant can vary across populations, places, and periods, and thus representation of the “natural environment” can change. For instance, Salganik et al.’s (2006) experimental music market may be more representative of today’s environment, with the ubiquity of online audio streaming services like Spotify and SoundCloud, than when the study was initially conducted. Thus, there is merit in sampling beyond what is considered representative of a specific population, place, time, or situation. Conditions change and can be changed; conditions that might not currently exist in the “natural environment” can still be the foundation for valuable technological and platform designs; for example, those that modify the “natural environment” of a digital social network platform to elicit desired behaviors from users.

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

R.2.3. Contemplating the nature of "Theory"

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

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

We empathize with these perspectives on the essential features of a theory. We acknowledge that theories can originate from various sources, including historical records, ethnographic observations, everyday anecdotal experiences, data mining, fitting a black box machine learning model, constructing analytically tractable mathematical models, or simply thinking hard about why people behave as they do. Nevertheless, a theory's truth is not guaranteed by satisfying any or all of these features, or by the method of its generation. It is also not clear why any specific feature or process of theorizing should always take precedence over others. Indeed, when pressed, researchers who advocate these features for theories will likely assert a connection between the presence of their preferred features

and the “success” of the theory in explaining existing observations and predicting unseen data. Hence, we view the following as the most inclusive criteria for what we expect from theories: they should (a) accurately explain observed experiment results and (b) make accurate predictions about unseen experiments. To be clear, we do not object to theories incorporating any other features. However, we argue that a theory’s primary evaluation criterion should be its accuracy in predicting unseen experimental outcomes, including out-of-distribution and under-intervention scenarios.

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

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

R.2.4. Diversity and Power Dynamics

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

In response to these concerns, we provide two clarifications. First, while we agree that diversity, equity, and inclusivity must be first-order considerations in any scientific reform, we see them as distinct issues from the intellectual merits of the argument for an integrative approach. Second, there are several mechanisms to ensure that a diversity of perspectives are heard. In the target article (see Sections 5.7 and 5.9) we highlighted the ways in which the integrative approach could allow a broader range of contributors to be involved in the process of designing, conducting, and analyzing experiments. The suggestions raised in the commentaries—such as adversarial collaborations and consensus meetings (see R.2.1)—provide further opportunities for participation.

In general, the responsibility of facilitating wide-ranging contributions to science is a burden shared by the entire scientific community. The integrative approach offers a distinctive set of opportunities and challenges in our collective pursuit of this goal: it expands the ways in which researchers can contribute to and benefit from research, but also potentially introduces new social dynamics we have to navigate to ensure this potential is achieved.

R3. Does the integrative approach have reach?

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

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

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

Lastly, **Haartsen et al.** highlight the potential value of the integrative approach in domains such as psychiatry and cognitive development. **Titone, Hernández-Rivera, Iniesta, Beatty-Martínez, and Gullifer** extend this approach to evaluate the implications of bilingualism on the mind and brain. They point to ongoing work consistent with the integrative approach, drawing connections between the *Systems Framework of Bilingualism* and research cartography. **Dohrn and Mezzadri** extrapolate the integrative approach to thought experiments. While these commentaries are clear and persuasive, our lack of expertise in these domains prevents us from contributing further substance to these discussions. The message from these commentaries is that issues stemming from the incommensurability characteristic of the "one-at-a-time" experiment design extend beyond the social and behavioral sciences and that the integrative approach may be fruitful in these domains and others beyond them.

Overall, these discussions and proposals sketch a vibrant picture of the various ways the integrative approach can be applied and expanded across different contexts. They reinforce

the value and potential of this methodology, offering much to consider and incorporate into our own research.

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

References

DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating Psychological Science With Metastudies: A Demonstration Using the Risky-Choice Framing Effect. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 17456916221079611.

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

Gelman, A. (2011). Causality and Statistical Learning. *The American Journal of Sociology*, *117*(3), 955–966.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Huang, L. (2023). A quasi-comprehensive exploration of the mechanisms of spatial working memory. *Nature Human Behaviour*, *7*(5), 729–739.

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

Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). *Exploring Limits to Prediction in Complex Social Systems* (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.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115.

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Musslick, S., Hewson, J., Andrew, B., Strittmatter, Y., Williams, C. C., Dang, G., Dubova, M., Holland, J. G. (in press). *An evaluation of experimental sampling strategies for autonomous empirical research in cognitive science*. In Proceedings of the 45th Annual Conference of the Cognitive Science Society. Sydney, AU.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*(5762), 854–856.

Suchow, J. W. (2023). Scaling up behavioural studies of visual memory [Review of *Scaling up behavioural studies of visual memory*]. *Nature Human Behaviour*, *7*(5), 672–673.

Watts, D. J., Beck, E. D., Bienenstock, E. J., Bowers, J., Frank, A., 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