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

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

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

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

Competing interest. None.

References

Awad, E., Dsouza, S., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM*, *63*(3), 48–55.

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

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Benitez-Quiroz, C. F., Srinivasan, R., & Martinez, A. M. (2016). EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In 2016 IEEE Conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA (pp. 5562–5570). https://doi.org/10.1109/CVPR.2016.600

Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2021). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, *54*(4), 1889–1901. https://doi.org/10.3758/s13428-021-01705-3

Hilton, C., & Mehr, S. (2021). Citizen science can help to alleviate the generalizability crisis.*45*, e21.

Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., & Mehr, S. A. (2023). How games can make behavioural science better. *Nature*, *613*(7944), 433–436.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Schutz, M., & Gillard, J. (2020). On the generalization of tones: A detailed exploration of non-speech auditory perception stimuli. *Scientific Reports*, *10*(1), 9520. https://doi.org/10.1038/s41598-020-63132-2

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1791.

Srinivasan, R., & Martinez, A. M. (2018). Cross-cultural and cultural-specific production and perception of facial expressions of emotion in the wild. *IEEE Transactions on Affective Computing*, *12*(3), 707–721.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001

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

Against naïve induction from experimental data

David Kellen[a]*, Gregory E. Cox[b], Chris Donkin[c], John C. Dunn[d] and Richard M. Shiffrin[e]

[a]Department of Psychology, Syracuse University, Syracuse, NY, USA; [b]Department of Psychology, College of Arts and Sciences, University at Albany, State University of New York, Albany, NY, USA; [c]Department of Psychology, Ludwig Maximilian University of Munich, München, Germany; [d]Department of Psychology, University of Western Australia, Perth, WA, Australia and [e]Psychological and Brain Sciences Department, Indiana University, Bloomington, IN, USA
davekellen@gmail.com
gecox@albany.edu
christopher.donkin@gmail.com
john.dunn@uwa.edu.au
shiffrin@indiana.edu

*Corresponding author.

doi:10.1017/S0140525X2300211X, e51

Abstract

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

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

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

https://doi.org/10.1017/S0140525X22002874 Published online by Cambridge University Press

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

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

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

(1) By placing a premium on commensurability, IED decreases the chances of new and unexpected findings (Shiffrin, Börner, & Stigler, 2018).
(2) By shifting researchers' resources toward the joint observation of a large number of factors, IED disrupts the piecemeal efforts in experimentation and theorization that illuminate the processes underlying human data generation. For instance, it makes it difficult to tell an important result from one caused by a confound (for discussions, see Garcia-Marques & Ferreira, 2011; Kellen, 2019; Shiffrin & Nobel, 1997).
(3) IED turns existential-abductive reasoning on its head: Instead of developing explanatory constructs (e.g., model development) in response to existing covariational information, a construct would be assumed a priori in the form of an empty vessel, to be later infused by the results of an experiment manipulating factors presumably related to it. For instance, the construct "attention" would be identified with the experimental manipulations thought to be relevant to "attention." This concern is materialized by the treatment of the so-called Moral Machine, a statistical model summarizing the observed relationships between moral judgments and a host of variables, as a bona fide theory of moral reasoning.
(4) By introducing a large number of factors, IED can easily degrade researchers' ability to identify which theoretical components are doing the leg work and which ones are failing, especially when compared to piecemeal testing (e.g., Birnbaum, 2008; Dunn & Rao, 2019; Kellen, Steiner, Davis-Stober, & Pappas, 2020). The recent application of IED to risky-choice modeling (Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021) illustrates this concern, as it is unclear which specific circumstances are leading one choice model to outperform another (e.g., is context dependency driven by feedback?).

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

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

Competing interest. None.

References

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

https://doi.org/10.1017/S0140525X22002874 Published online by Cambridge University Press

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

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

Erich Kummerfeld[a]* and Bryan Andrews[b]

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

*Corresponding author.

doi:10.1017/S0140525X23002273, e52

Abstract

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

Almaatouq et al. propose a break from tradition to accelerate scientific progress, and we applaud them for it. However, we urge an even further shift to incorporate theory and methods from causal discovery, a subfield of machine learning with decades of research on artificial intelligence (AI)-guided causal learning and experiment design. Causal discovery has not been well leveraged in the experimental sciences perhaps because it also breaks from tradition – statistical tradition.

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

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

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

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

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

https://doi.org/10.1017/S0140525X22002874 Published online by Cambridge University Press