Commensurability engineering is first and foremost a theoretical exercise

Joachim Vandekerckhove[a,*]

Author final version. This is a comment on *Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences* by Abdullah Almaatouq, Thomas L. Griffiths, Jordan W. Suchow, Mark E. Whiting, James Evans, and Duncan J. Watts, to appear in *Behavioral and Brain Sciences*.

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

Commensurability engineering | Metastudies | Mathematical Social Science

When we first executed a meta-study in 2015 (Baribault, 2019; Baribault et al., 2018), the concept of sampling from a method space (what the target article calls a "design space") was central to its implementation. We had set out to replicate an interesting effect we had found in a published paper. However, we soon realized that we would need to specify so many details of implementation—the kinds of things researchers rarely make explicit in their methods sections—that we felt we could not perform a faithful replication. Of course, we could have reached out to the original authors, but we also felt that the literature should to some extent be able to stand on its own. Eventually, we decided to be good Bayesians and allow for uncertainty in our experimental design. In contrast to a "point experiment," a meta-study defines a distribution over the method space, from which we can draw samples in a kind of Monte Carlo integration over our uncertainty as to which point experiment best captures the effect of interest.

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

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

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

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

An underappreciated strength of meta-studies is their statistical efficiency (DeKay, Rubinchik, Li, & De Boeck, 2022; Rubinchik, 2019). In a meta-study, increasing the number of point experiments k reduces the standard error of the mean effect size above and beyond the total number of participants P. To see this, consider the equation for the error variance in a random-effects meta-analysis as a function of the variance in effect sizes across subjects σ^2 and the variance in effect sizes across studies τ^2:

$$\varsigma^2 = \frac{\sigma^2}{P} + \frac{\tau^2}{k}.$$

For a fixed number of participants, increasing the number of point experiments (and reducing the number of participants per study) maximizes estimation accuracy.

[a]University of California, Irvine
[*]Correspondence concerning this article should be addressed to Joachim Vandekerckhove (joachim@uci.edu).
This work was supported by National Science Foundation grants #1754205, #1850849, and #2051186.

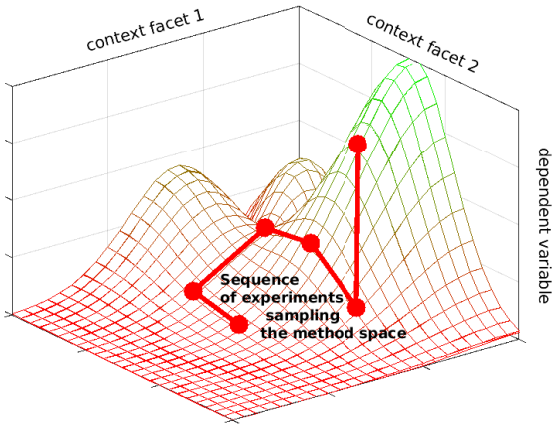

Fig. 1. A guided walk through a two-dimensional method space, finding a configuration of facets that optimizes the dependent variable. The DV may be as simple as an effect size or as sophisticated as a Bayes factor. This figure does not appear in the comment submitted to *Behavioral and Brain Sciences*.

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

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

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

References

Baribault, B. (2019). *Using hierarchical Bayesian models to test complex theories about the nature of latent cognitive processes.* Retrieved from https://worldcat.org/title/1114623187 (Unpublished doctoral dissertation.)

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., ... Vandekerckhove, J. (2018, March). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612. doi: 10.1073/pnas.1708285114

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163.

DeKay, M. L., Rubinchik, N., Li, Z., & De Boeck, P. (2022). Accelerating psychological science with metastudies: A demonstration using the risky-choice framing effect. *Perspectives on Psychological Science*, *17*(6), 1704–1736.

Kenett, R. S., & Rubinstein, A. (2021). Generalizing research findings for enhanced reproducibility: an approach based on verbal alternative representations. *Scientometrics*, *126*(5), 4137–4151.

Rubinchik, N. (2019). *A demonstration of the meta-studies methodology using the risky-choice framing effect.* Retrieved from https://worldcat.org/title/1159175086 (Unpublished Master's thesis.)

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behavior*, *1*(1), 0015.