Rapid #: -19137870

We need to be braver about the generalizability crisis

Published online by Cambridge University Press: **10 February 2022**

Todd S. Braver (iD) and Sanford L. Braver

Commentary Related commentaries Metrics

Behavioral and Brain Sciences

Article contents

Abstract

We applaud the effort to draw attention to generalizability concerns in twenty-first-century psychological research. Yet we do not feel that a pessimistic perspective is warranted. We outline a continuum of available methodological tools and perspectives, including incremental steps and meta-analytic approaches that can be readily and easily deployed by researchers to advance generalizability claims in a forward-looking manner.

We heartily applaud and commend Yarkoni for drawing attention to issues of generalizability in twenty-first-century psychological science. We strongly concur that these issues are of crucial importance; nevertheless, at least for most cognitive scientists and neuroscientists, they have not been prominent (in contrast to the voluminous literature

in field and applied research, under the heading of *external validity*; e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979). Moreover, we agree with Yarkoni's perspective that the current preoccupation with reproducibility and replication may be misplaced, given that generalizability is a logically prior and potentially stronger concern. However, we definitely do not share Yarkoni's pessimistic perspective. Certainly, we do not support the extrapolation from this perspective to suggestions that academic psychologists consider pursuing different careers or switching from quantitative to qualitative research. Instead, we contend that there are ripe opportunities for psychological researchers to advance the generalizability of key phenomena of interest, by making greater use of the full continuum of available methods that can be deployed for this purpose.

The *radical randomization (RR)* experiment (Baribault et al., 2018; highlighted in Yarkoni) anchors one pole of this continuum, as the most ambitious and comprehensive strategy. An RR experiment involves *many* (16 in Baribault et al., 2018) potentially irrelevant factors – or moderators – that are varied *randomly* within the experimental design (as "micro-experiments"). With Bayesian hierarchical modeling, both the summary effect size and the moderating effect of each random factor can be properly estimated. However, as a highly effort- and resource-intensive endeavor, the RR experiment seems less likely to serve as the primary approach for addressing generalizability concerns.

Fortunately, more easily deployed approaches are available. We agree with Shadish, Cook, and Campbell (2002), who eschew both of the key defining features (italicized above) of RR studies: (1) that researchers simultaneously address multiple potential (often theoretically irrelevant) moderators at once in the same meta-study; and (2) that the levels of these moderating factors be both randomly selected and analyzed as random- (rather than fixed-) factors. The objection to (1) is based on the insight that there will always remain a virtually infi

space of additional, possible, non-varied factors that could limit generalizability. Indeed, the very nature of inductive logic precludes ever completely resolving generalizability issues. Nevertheless, some inferential purchase can be provided – albeit somewhat more slowly – via an *incremental* (i.e., study-by-study), rather than comprehensive, strategy. With respect to (2), although random selection and random-effects models are clearly preferred, researchers can still legitimately advance the generalizability of their postulates by "guessing at laws and checking out some of these generalizations in other *equally specific* but *different* conditions" (Campbell & Stanley, 1963, p. 17, emphasis added).

Thus, to anchor the other pole of generalizability efforts, we propose that researchers consider varying at least one, unique, and supposedly irrelevant contextual factor in each experiment (see Yarkoni, p.9 for examples). Importantly, even with only a few (but of course more than one) *purposively* selected levels of this factor, there is still an interpretational advantage to be gained. Specifically, even if using fixed-effects rather than random-effects analysis, the *interaction* of this factor with the main effect of interest can be tested, to estimate its impact. Only if the interaction effect is small and insignificant can the claim be made that the induced heterogeneity is indeed plausibly irrelevant; if so, generalizability claims over this factor can be furthered. For example, imagine if, in the original Schooler and Engstler-Schooler (1990) study highlighted by Yarkoni, multiple perpetrator videos had been used, with similar effect sizes for each (i.e., no interaction). Moreover, this approach enables generalizability claims regarding a phenomenon of interest to be advanced incrementally, study-by-study. Generalizability claims become more grounded and justifiable – albeit with greater effort – by moving along the continuum toward RR: varying additional putative nuisance factors, including more exemplars of each factor, selecting (sampling) these exemplars at random rather than purposively, and evaluating them with random-effects rather than fixed-effects analyses.

The arguably mid-continuum issue of conceptual, as opposed to exact, replications also highlights our key disagreement with Yarkoni. In particular, we claim that Yarkoni unfairly undersells the substantial epistemological advantage of conceptual replications. Of course, he is not alone: researchers often implement replications by trying to precisely match all the details of an initial study, presumably out of a superstitious desire to "get it right." Yet, as many have long pointed out (Brunswik, 1956; Campbell & Stanley, 1963; Cronbach, 1975), the stronger alternative is to purposely vary the features that should be theoretically irrelevant, with the goal of finding that such variation does not in fact alter the outcome. Yarkoni dismisses conceptual replications, by alleging that they "do not lend themselves well to a coherent modeling strategy. . . . It is rarely obvious how one can combine the results to obtain a meaningful estimate of the robustness or generalizability of the common effect (p.25)."

We strongly disagree with Yarkoni on this critical point. In particular, meta-analysis techniques are precisely designed to evaluate the robustness of effect size findings over a set of studies. In addition to statistics that quantify summary (overall) effect size, it is standard to also evaluate homogeneity of effect, i.e., through indices as the Q-test and the I^2 statistics (Hedges & Olkin, 1985). Meta-analysis is typically invoked for retrospective reviews of a body of literature, yet Braver, Thoemmes, and Rosenthal (2014) extend its utility via the *continuously cumulating meta-analytic* approach. With this approach, meta-analytic calculations can be employed incrementally – even within-study (i.e., across experiments) – as new findings emerge. Newer Bayes Factor approaches may gain even greater traction as a means of directly implementing this incremental perspective (Scheibehenne, Jamil, & Wagenmakers, 2016).

In summary, our goal is to help psychological researchers appreciate that there is an entire buffet of experimental design and analysis options readily available and waiting to be deployed to address

issues of generalizability. There is no need to despair, or begin searching out alternative career choices! Indeed, all that is needed is for the field to face the generalizability crisis in a *Braver* manner.

Financial support

Conflict of interest

None.

References

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. Proceedings of the National Academy of Sciences, 115(11), 201708285. https://doi.org/10.1073/pnas.1708285114. CrossRef Google Scholar PubMed GET IT!

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. Perspectives on Psychological Science, 9(3), 333–342. https://doi.org/10.1177/1745691614529796. CrossRef Google Scholar PubMed GET IT!

Brunswik, E. (1956). Perception and the representative design of psychological experiments (2nd ed.). University of California Press. CrossRef Google Scholar GET IT!

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Rand-McNally. Google Scholar GET IT!

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation:

Design and analysis issues for field settings. Rand-McNally. Google Scholar GET IT!

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30(2), 116–127. https://doi.org/https://psycnet-apa-org.libproxy.wustl.edu/doi/10.1037/h0076829. CrossRef Google Scholar GET IT!

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press. https://doi.org/https://doi.org/10.1016/C2009-0-03396-0. Google Scholar GET IT!

Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results. Psychological Science, 27(7), 1043–1046. https://doi.org/10.1177/0956797616644081. CrossRef Google Scholar PubMed GET IT!

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 22(1), 36–71. https://doi.org/10.1016/0010-0285(90)90003-m. CrossRef Google Scholar PubMed GET IT!

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin and Company. Google Scholar GET IT!

✅ Access

In response to: **The generalizability crisis**

Related commentaries (38) Author response

Related content

AI-generated results: by **UNSILO**

 Article

Predicted effect size of lisdexamfetamine treatment of attention deficit/hyperactivity disorder (ADHD) in European adults: Estimates based on indirect analysis using a systematic review and meta-regression analysis

M. Fridman, P.S. Hodgkins, J.S. Kahle and M.H. Erder

European Psychiatry

Published online: 15 April 2020

Article

Is there an excess of significant findings in published studies of psychotherapy for depression?

J. Flint, P. Cuijpers, J. Horder, S. L. Koole and M. R. Munafò

Psychological Medicine

Published online: 25 July 2014

Article

Executive Function in Pediatric Sleep-Disordered Breathing: A Meta-analysis

Jonathan J. Mietchen, David P. Bennett, Trevor Huff, Dawson W. Hedges and Shawn D. Gale

Journal of the International Neuropsychological Society

Published online: 2 August 2016

Article

Précis of _Statistical significance: Rationale, validity, and utility_

Siu L. Chow

Behavioral and Brain Sciences

Published online: 1 April 1998

Article

Comparison of psychotherapies for adult depression to pill placebo control groups: a meta-analysis

P. Cuijpers, E. H. Turner, D. C. Mohr, S. G. Hofmann, G. Andersson, M. Berking and J. Coyne

Psychological Medicine

Published online: 3 April 2013

Element

Confronting School Violence

Jillian J. Turanovic, Travis C. Pratt, Teresa C. Kulig and Francis T. Cullen

Published online: 18 March 2022

Article

The Benefit of Interventions to Reduce Posttraumatic Stress in Youth Exposed to Mass Trauma: A Review and Meta-Analysis

Betty Pfefferbaum, Pascal Nitiéma, Elana Newman and Anushka Patel

Prehospital and Disaster Medicine

Published online: 28 August 2019

Article

A meta-analysis of the neuropsychological sequelae of HIV infection

MARK REGER, ROBERT WELSH, JILL RAZANI, DAVID J. MARTIN and KYLE B. BOONE

Journal of the International Neuropsychological Society

Published online: 16 May 2002