frontiers | Frontiers in Organizational Psychology

TYPE Field Grand Challenge
PUBLISHED 04 October 2023
DOI 10.3389/forgp.2023.1237756

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Darren C. Treadway,
Niagara University, United States

*CORRESPONDENCE
Piers Steel
✉ steel@ucalgary.ca

RECEIVED 09 June 2023
ACCEPTED 15 September 2023
PUBLISHED 04 October 2023

CITATION
Steel P (2023) Leftovers and boundary conditions: a moderator proposal. *Front. Organ. Psychol.* 1:1237756. doi: 10.3389/forgp.2023.1237756

Leftovers and boundary conditions: a moderator proposal

Piers Steel*

Department of Organizational Behaviour and Human Resources, Haskayne School of Business, University of Calgary, Calgary, AB, Canada

KEYWORDS

generalizability, moderators, boundary conditions, meta-analysis, Problem of Induction

Systematic reviews and meta-analyses are a chance for a field to take stock and see what is now known (Hunt, 1997). Articles are searched for, then gathered, and then coded into a common metric, allowing for statistical analysis. Inevitably, for any meta-analysis of a decent size, the effect sizes from all these studies are not the same: variation around the mean is the norm (Steel et al., 2015). Part of this is due to sampling error, where the random chance draw of participants or data points accounts for fluctuations. However, even after accounting for the uncertainty associated with finite sample size, there is residual variance—the leftovers. Like culinary leftovers, they can seem like an afterthought, and some have argued they are of no great importance (LeBreton et al., 2017). However, accounting for them is foundational to the advancement of our science.

This leftover variance goes by a variety of names, from tau to the REVC (Random Effects Variance Component). It reflects the possible outcomes or the sizes of your effect sizes, the range of which is called credibility or prediction intervals. Basically, if a study was redone within the confines of what was done before, it effects size should be within these intervals. Credibility intervals are often broad and can cross the correlational Rubicon of zero, where effect sizes fail to even directionally generalize. Accounting for this variation can be all too important. Without it, each of our studies is simply a snapshot frozen in time, speaking to what happened in that particular moment in that specific setting, which may or may not happen again or at least not to the same extent (Yarkoni, 2022). On the other hand, if we can identify the sources of variation, we have a pathway to a mature science that can make precise predictions based on diagnosis or assessment alone. It can tell you when and where a finding is applicable and to who. For example, a medical treatment might cure some and kill others, so best if we could predict that variation. Unfortunately, we often can't. In the sobering words of Flake et al. (2022): "any statistical model estimated from any study has so many omitted sources of variance that the estimates are likely meaningless" (p. 33).

When a meta-analysis is conducted, a moderator analysis tries to account for this variation in effect sizes. Along these lines, we have advanced statistical methodology quite far, moving from simply subgroup analysis, where we compare the effect sizes of two group and see if one is bigger, to souped-up multiple regression schemes using continuous variables and sophisticated weighting (Steel et al., 2021). Despite these statistical refinements, often meta-analysts find that the moderators they want are not in the literature obtained (Steel et al., 2015). Aside from type of measure, studies typically confine their reporting to the thin gruel of participant age, gender ratio, student status, and nation. If you see an abundance of meta-analyses that use culture as a moderator, well there is a reason. Despite meta-analyses being invaluable summaries, and often used for precisely that, journals insist on new theory being tested. With a paucity of moderators available, they can link nation to national culture and that comes with the requisite theory. Still, the end result is that we still largely do not know what moderates relationships.

Consequently, we as a field are in a bind. Our studies do not properly contextualize themselves and our sample descriptions, comprising mostly of easily obtained demographics, have become a mostly empty ritual. Often, they are not the moderators we are looking for. Furthermore, it is not even clear how a study should be contextualized, how

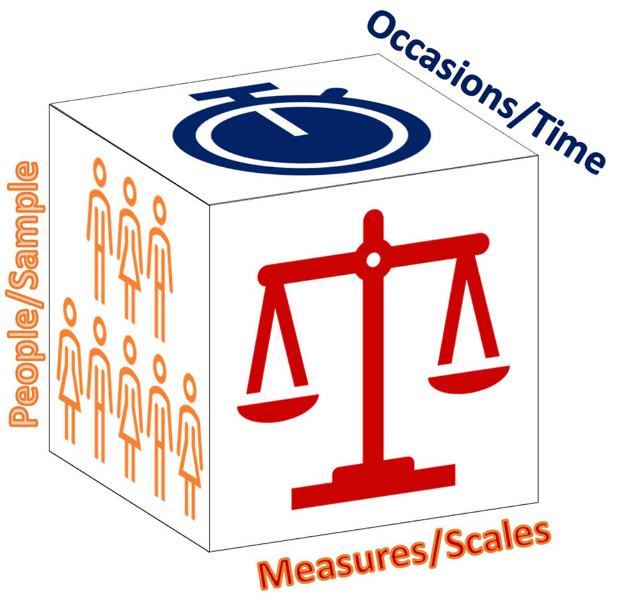

FIGURE 1
Cattell's Data Box/Cube.

it should be described. To advance as a field, we need to know our boundary conditions, that is when an effect or a conclusion will hold. This is our Grand Challenge.

There are a variety of ways of establishing of boundary conditions, as per the 38 commentaries associated with Yarkoni's (2022) review "The Generalizability Crisis". Qualitative review articles on established moderators and their expected impact are invaluable, especially if accompanied by sufficient theoretical explanation. Ideally, this would lead to exact recommendations of how studies should be described. For example, if studies routinely linked employees' jobs to their O*NET occupational codes and their organization's Standard Industrial Classification Codes or equivalent (even in a supplementary file), this would be tremendous boon to the field. Also, social or technological forecasting proposals or papers regarding how we could account for the variation or what could be achieved once we have prediction would provide legitimacy and motivation for the endeavor. These can include efforts toward Community Augmented Meta-Analysis (CAMAs) or mass replication efforts (e.g., Visser et al., 2022). For example, combining moderator search with meta-analytic structural equation modeling (MASEM) allows us to predict validity coefficients without a local validation study and build instantly personnel selection systems that are costless and of high quality (Steel et al., 2010).

Finally, as fodder for our former efforts, empirical articles testing when differences matter would be invaluable. This can range from field experiments to surveys. As for their focus, there are a variety of decompositions of these moderators, from Cattell's (1966) ten-dimensional system to simpler schemes used to organize meta-analyses (e.g., PICO, SPIDER, SPICE). To highlight possible contributions, the three fundamental dimensions of Cattell's Data Box/Cube will suffice (Revelle, 2009): People, Measurements, and Occasions (see Figure 1).

People

When does the specifics of a sample make a difference? It can of course. Relatively recently, heart disease in women was more frequently mis-diagnosed than men as research favored male only samples and women's heart attacks can present differently (Lancet, 2019). The decades it took to establish this highlights our need. In the other direction, we often hypothesize differences where there may be none and act as it has already been proven. As an example of exemplary branding, we have Western, Educated, Industrialized, Rich and Democratic nations or WEIRD (Henrich et al., 2010). It emphasizes that historically a preponderance of research was conducted in the United States, and due to cultural differences, results *may* not generalize. Or they might. In general, theories of cultural determinism encounter the core problem that the variation within cultures overwhelmingly swamps the variation between. In other words, we shouldn't use "country averages to act as proxies for cultural values of individuals or small groups from these countries" (Taras et al., 2016, p. 483). It helps explain why across 125 samples testing 28 classic findings there was little evidence of WEIRD cultures being a moderator (Klein et al., 2018). This isn't to say that cultural values aren't important for national lines of research (Steel et al., 2018), but we have limited insight regarding when national values are relevant for individual and group behavior (Schimmelpfennig et al., under review)[1].

Specifically for the field of organizational behavior, we have our own bias against using student samples in lieu of full-time employees. In fact, many business-themed journals have warning against substitution (e.g., Bello et al., 2009), justified as a threat to validity, which may or may not manifest. For example, we could easily have two identical groups, but described in different terms. There are employees who are upgrading their education as MBA students and there are MBA students who are paying for classes by working full-time. Characterize your sample as the former and it is publishable in many journals but not as the latter. Alternatively, we could sample students 1 week prior to graduation or a week after employment. During that brief hiatus, have there really been substantive changes? Furthermore, students are hardly a monolithic group and differences among them can be due to tuition cost, degree program, and type of degree, just as there can be socio-economic, occupational, and experiential effects. As Taras et al. (2023) conclude, "Rather than automatically rejecting student samples, we should take a more nuanced approach and establish the boundary conditions of their generalizability, a long, long overdue task" (p. 10).

Measurements

The jingle-jangle or commensurability problem is endemic. We have incompatible measures that go by the same name (i.e., jingle)

[1] Schimmelpfennig, R., Spicer, R., White, C. J., Gervais, W., Norenzayan, A., Heine, S., et al. (under review). A problem in theory and more: measuring the moderating role of culture in Many Labs 2. *PsyArXiv*. doi: 10.31234/osf.io/hmnrx

and functionally identical ones that go by different names (i.e., jangle). Reviewed by Bosco et al. (2015) under the heading "Tower of Babel," this is major impediment to scientific advancement, hampering efforts toward functionally taxonomies. For example, Taras et al. (2023) examined seven cultural value scale, finding that results can completely invert, switch from positive to negative, for a variety of findings depending on the scale used. After sampling error, method variance is typically the largest contributor to fluctuations in effect sizes (Kammeyer-Mueller et al., 2010).

This problem of construct and measurement proliferation will only intensify as there are considerable incentives to creating new measures. Consider the popular success of the GRIT measure, which is essentially a conscientiousness clone (Credé et al., 2017; King and Wright, 2022). Ideally, cleaning up constructs and their associated measures should exceed the pace of creating them, which is possible. Meta-analyses or single studies can tackle multiple measures at the same time, showing what is compatible and what should be kept separate. Validation efforts can be shared among those functionally identical. For example, Park et al. (2020) considered differences among personality measures and Richard et al. (2009) sorted indices of organizational performance. There is also opportunity for consolidation. As Taras et al. (2023) review, researchers can pick and choose individual items from multiple scales to create the best overall version. These efforts should culminate into taxonomic work that organizes a broad range of constructs. Presently, as reviewed by Steel et al. (2021) "The multiplicity of overlapping terms and measures creates a knowledge management problem that is increasingly intractable for the individual researcher to solve" (p. 30).

Occasions

Occasions or conditions can be seen as the context under which the study was conducted. It is essentially why field experiments' results differ from those conducted in the lab. In the laboratory, we have control and can eliminate sources of "error" (i.e., influential aspects that are not the focus of study) but field experiments, being immersed in the extraneous, provide realism and can confirm the robustness of an effect in a complex and intertwined world. Given all that can occur contextually, this is the broadest of the categories and partially overlaps with the former two. The nature of demographic groups can change over time, such as generational effects, though they are often confused with maturation (Steel and Kammeyer-Mueller, 2015). Also, the meaning or interpretation of a scale can shift. The original Eysenck Personality Questionnaire, for example, would ask "Do you lock up your house carefully at night?" as an indicator of psychopathy (Eysenck et al., 1985), which may have been a reasonable indicator if it was the 1950s and you lived in a small town in Britain.

The most ubiquitous way of capturing occasions is time, which encapsulates zeitgeist's entirety. This is often studied under the terms of validity degradation (Keil and Cortina, 2001) and the decline effect (Schooler, 2011). That our findings age-out is such a common phenomenon so we should expect what was once a dependable relationship may be no more. We have a sufficient body of research that longitudinal investigation should be routine, often called cross-temporal meta-analysis (Rudolph et al., 2020).

Conclusion

For Grand Scientific Challenges, this issue of generalizability is core. Hume discussed it as the philosophical Problem of Induction, as per "There can be no demonstrative arguments to prove, that those instances, of which we have had no experience, resemble those, of which we have had experience" (Hume, 1739-1740/1888, p. 89). Contrast this with Matt and Cook (2009) grappling with the limitations of scientific conclusions, that is "the issue is how one can justify inferences to these novel universes of persons, treatments, outcomes, setting and times on the basis of findings in other universes" (p. 513). In short, how do we know what generalizes? Though there can be no definitive solution, we can make plausible steps toward shrinking our uncertainty. Afterall, for us to have an applied science, we need to understand when our findings apply.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bello, D., Leung, K., Radebaugh, L., Tung, R. L., and Van Witteloostuijn, A. (2009). From the editors: Student samples in international business research. *J. Int. Bus. Stud.* 40, 361–364. doi: 10.1057/jibs.2008.101

Bosco, F. A., Steel, P., Oswald, F. L., Uggerslev, K., and Field, J. G. (2015). Cloud-based meta-analysis to bridge science and practice: welcome to metaBUS. *Person. Assess. Decisions* 1.

Cattell, R. B. (1966). "The data box: Its ordering of total resources in terms of possible relational systems," in *Handbook of Multivariate Experimental Psychology,* ed. R. B. Cattell. Chicago: Rand-McNally, 67–128.

Credé, M., Tynan, M. C., and Harms, P. D. (2017). Much ado about grit: a meta-analytic synthesis of the grit literature. *J. Pers. Soc. Psychol.* 113, 492–511. doi: 10.1037/pspp0000102

Eysenck, S. B., Eysenck, H. J., and Barrett, P. (1985). A revised version of the psychoticism scale. *Pers. Individ. Dif.* 6, 21–29. doi: 10.1016/0191-8869(85)90026-1

Flake, J. K., Luong, R., and Shaw, M. (2022). Addressing a crisis of generalizability with large-scale construct validation. *Behav. Brain Sci.* 45, e14. doi: 10.1017/S0140525X21000376

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature* 466, 29–29. doi: 10.1038/466029a

Hume, D. (1739-1740/1888). *Hume's Treatise Of Human Nature.* Oxford, England: Clarendon Press.

Hunt, M. (1997). *How Science Takes Stock: The Story of Meta-Analysis.* New York City: Russell Sage Foundation.

Kammeyer-Mueller, J., Steel, P. D., and Rubenstein, A. (2010). The other side of method bias: the perils of distinct source research designs. *Multivar. Behav. Res.* 45, 294–321. doi: 10.1080/00273171003680278

Keil, C. T., and Cortina, J. M. (2001). Degradation of validity over time: a test and extension of Ackerman's model. *Psychol. Bull.* 127, 673. doi: 10.1037/0033-2909.127.5.673

King, K. M., and Wright, A. G. (2022). A crisis of generalizability or a crisis of constructs? *Behav. Brain Sci.* 45, e24. doi: 10.1017/S0140525X21000443

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., et al. (2018). Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* 1, 443–490. doi: 10.1177/2515245918810225

Lancet, T. (2019). Cardiology's problem women. *Lancet (London, England)* 393, 959. doi: 10.1016/S0140-6736(19)30510-0

LeBreton, J. M., Schoen, J. L., and James, L. R. (2017). "Situational specificity, validity generalization, and the future of psychometric meta-analysis," in *Handbook of Employee Selection*, eds. J. L. Farr & N. T. Tippins. New York, NY: Routledge, 93–114.

Matt, G. E., and Cook, T. D. (2009). "Threats to the validity of generalized inferences," in *The Handbook of Research Synthesis and Meta-Analysis*, eds. H. Cooper, L. V. Hedges, and J. C. Valentine. New York: Russell Sage Foundation, 537–560.

Park, H. H., Wiernik, B. M., Oh, I. S., Gonzalez-Mul,é, E., Ones, D. S., and Lee, Y. (2020). Meta-analytic five-factor model personality intercorrelations: Eeny, meeny, miney, moe, how, which, why, and where to go. *J. Appl. Psychol.* 105, 1490–1529. doi: 10.1037/apl0000476

Revelle, W. (2009). Personality structure and measurement: the contributions of Raymond Cattell. *Br. J. Psychology* 100, 253–257. doi: 10.1348/000712609X4 13809

Richard, P. J., Devinney, T. M., Yip, G. S., and Johnson, G. (2009). Measuring organizational performance: towards methodological best practice. *J. Manage.* 35, 718–804. doi: 10.1177/0149206308330560

Rudolph, C. W., Costanza, D. P., Wright, C., and Zacher, H. (2020). Cross-temporal meta-analysis: A conceptual and empirical critique. *J. Bus. Psychol.* 35, 733–750. doi: 10.1007/s10869-019-09659-2

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature* 470, 437–437. doi: 10.1038/470437a

Steel, P., Beugelsdijk, S., and Aguinis, H. (2021). The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews. *J. Int. Bus. Stud.* 52, 23–44. doi: 10.1057/s41267-020-00385-z

Steel, P., Johnson, J., Jeanneret, P. R., Hoffman, C., Scherbaum, C., and Foster, J. (2010). At sea with synthetic validity. *Ind. Organ. Psychol.* 3, 371–383. doi: 10.1111/j.1754-9434.2010.01255.x

Steel, P., and Kammeyer-Mueller, J. (2015). The world is going to hell, the young no longer respect their elders, and other tricks of the mind. *Ind. Organ. Psychol.* 8, 366–371. doi: 10.1017/iop.2015.51

Steel, P., Kammeyer-Mueller, J., and Paterson, T. A. (2015). Improving the meta-analytic assessment of effect size variance with an informed Bayesian prior. *J. Manage.* 41, 718–743. doi: 10.1177/0149206314551964

Steel, P., Taras, V., Uggerslev, K., and Bosco, F. (2018). The happy culture: a theoretical, meta-analytic, and empirical review of the relationship between culture and wealth and subjective well-being. *Pers. Soc. Psychol. Rev.* 22, 128–169. doi: 10.1177/1088868317721372

Taras, V., Steel, P., and Kirkman, B. L. (2016). Does country equate with culture? Beyond geography in the search for cultural boundaries. *Manag. Int. Rev.* 56, 455–487. doi: 10.1007/s11575-016-0283-x

Taras, V., Steel, P., and Stackhouse, M. (2023). A comparative evaluation of seven instruments for measuring values comprising Hofstede's model of culture. *J. World Business.* 58, 101386. doi: 10.1016/j.jwb.2022. 101386

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., et al. (2022). Improving the generalizability of infant psychological research: the ManyBabies model. *Behav. Brain Sci.* 45, e1. doi: 10.1017/S0140525X210 00455

Yarkoni, T. (2022). The generalizability crisis. *Behav. Brain Sci.* 45, e1. doi: 10.1017/S0140525X20001685