Replication versus Generalization: A Comparison of two Frames for Understanding Building Evidence for Program Effectiveness

Mark White
University of Oslo

Author note

Correspondence concerning this abstract should be addressed to Mark White,

Frognerveien 57, Oslo 0266. Norway. E-mail: mrkwht@umich.com

Background/Purpose:

It is vital that we know if research findings are true. Replication is the dominant paradigm behind this goal. I use the word paradigm because while disagreement around what specifically counts as a replication exists (see Bollen, et al., 2015), replication as a way of determining scientific truth is known by all: one study estimates a program's effectiveness and future studies confirm the finding, ensuring it effect is real. In this paper, I argue that the replication paradigm is an important cognitive frame that promotes counter-productive understandings of evidence in education. Instead, we should adopt a frame of generalization.

Methodology:

This conceptual paper uses frame analysis (Goffman, 1974), which involves analyzing how words and phrases guide specific forms of sense-making and constrain the way we understand the world, often through connecting with deeper concepts. Starting with education articles returned by google scholar with the word "replication" in the title and expanding through mining reference lists, I examined the way replication was discussed in the literature, looking for implicit assumptions inherent in the use of the concept. I note here that many of those on the cutting edge of replication research adopt points of view much like the generalization frame that I will promote and I focus here on ways of thinking by those not deeply involved in thinking about replication.

Findings/Results:

The Replication Frame: Replication focuses on the question, 'is this really true?', often based on concerns arising from measurement or sampling errors. Some, though not all (e.g. Asendorpf, et al., 2013; Bollen, et al., 2015), have expanded replication to focus on the question "is it true in some other context?" (Schmidt, 2009; where context can be defined very broadly). In either case, replication seems to imply a number of assumptions that are problematic[1].

The first problematic assumption is that replications should be successful. Well-conducted, well-powered studies control for errors and thus *should* replicate. This is in fact not true, but a common misinterpretation of p-values (Kline, 2004); a significant finding does not tell one how likely the effect is to be replicated. This assumption leads to the undervaluing of replications, because it implies replications are not novel, but simply verify known facts. It also implies the "replication crisis". The replication crisis is a *crisis* because it violated the implicit assumption that effects should replicate, which calls into doubt the foundation of the replication frame.

The second problematic assumption is that it encourages the implicit extrapolation from studies to programs. An experiment measures the effect of a program against some comparison. Especially when the experiment is replicated, and so we validate the claim that the program is effective, this promotes the view that the program itself is now evidence-based, a view which has been codified into law through ESSA (see Slavin, 2017). However, generalizing from the effect

[1] The careful reader might notice that the replication paradigm as I lay it out is intimately tied to null-hypothesis statistical testing (NHST), which is becoming increasingly controversial (Kline, 2004). In fact, we could probably replace the "replication paradigm/frame" with the NHST frame.

of a few specific studies to make a conclusion about a program is an incredible feat of extrapolation in most cases, as even a series of replicated studies examines a program in a limited set of contexts.

Last, the replication frame implies that replicated effects are more trustworthy. This is not always the case. The likelihood that an effect is actually true depends on false-positive rates, power, and the a prior likelihood that the effect is true (Ioannidis, 2005). One large, well-conducted study can provide a greater post-study likelihood that an effect is true than multiple smaller, poorly run studies.

Generalizability Theory as an Alternative Frame

Generalizability theory (GTheory; Brennan, 2001) provide a better frame for building scientific evidence. GTheory maintains that any measurement (such as the effect estimated by an experiment) is the result of many specific facets, where facets are contextual features of the measurement process that affect the measured value. Measurements are assumed bound within the measurement context until evidence is presented to suggest otherwise. Applied to research, this acknowledges that a study occurs with specific *schools* and specific *students* at a specific *time,* is conducted by specific *researchers*, and uses specific *outcomes* and *analytic approaches*. GTheory provides a framework for understanding the extent to which these contextually bound measurements dependably (i.e. reliably) generalize to a pre-specified broader universe. Rather than asking "is it true?"; GTheory asks a combination of "how large is the effect? How dependent is it on specific facets?". It flips the assumptions, requiring evidence to prove generalization rather than implicitly assuming generalization should occur. This removes the problematic assumptions implicit in the replication frame, providing a frame to understanding how to build research evidence while focusing attention on where effects might apply, what factors might moderate effects, and how effects vary across study facets.

Implications:

Adopting a generalizability frame would require revamping both policies around evidence-based programs and ways that program effectiveness data is archived. For example, the ESSA evidence standards and the What Works Clearinghouse (WWC), rely on designating programs as evidence-based, which does not fit the GTheory frame. I focus here on suggested changes to WWC as a way of highlighting the implications. First, at the very least, WWC could add badges or notices that indicate when evidence on a program comes from a limited source, such as program developers[2]. Second, rather than general designations of being evidence-based, practitioners should be able to enter their school or district's characteristics and get a customized prediction for how effective the program would be in their context (with some uncertainty measure), including an acknowledgment that not enough evidence exists when that is the case. This would require experimental studies to archive school specific (anonymized but with demographic characteristics) estimates of program effectiveness to get school specific estimates. This is a major overhaul, but abandons problematic assumptions inherent in the replication frame.

[2] This has been done in a limited fashion for school demographics.

References

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318–335.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. National Science Foundation.

Brennan, R. L. (2001). *Generalizability Theory*. Retrieved from http://link.springer.com/10.1007/978-1-4757-3456-0

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA, US: Harvard University Press.

Kline, R. (2004). What's Wrong with Statistical Tests- And Where we go from here. In *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research*. Retrieved from https://apastyle.apa.org/manual/related/kline-2004.pdf

Slavin, R. E. (2017). Evidence-Based Reform in Education. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(3), 178–184. https://doi.org/10.1080/10824669.2017.1334560