Journal of Social and Political Psychology
jspp.psychopen.eu | 2195-3325

Commentaries

The Importance of Heterogeneity in Large-Scale Replications

Matthew H. Goldberg*[a], Sander van der Linden[b]

[a] Yale Program on Climate Change Communication, Yale University, New Haven, CT, USA. **[b]** Department of Psychology, University of Cambridge, Cambridge, United Kingdom.

Abstract

In a large-scale replication effort, Klein et al. (2018, https://doi.org/10.1177/2515245918810225) investigate the variation in replicability and effect size across many different samples and settings. The authors concluded that, for any given effect being studied, heterogeneity across samples and settings does not explain failures to replicate. In the current commentary, we argue that the heterogeneity observed indeed has implications for replication failures, as well as for statistical power and theory development. We argue that psychological scientific research questions should be contextualized—considering how historical, political, or cultural circumstances might affect study results. We discuss how a perspectivist approach to psychological science is a fruitful way for designing research that aims to explain effect size heterogeneity.

Keywords: heterogeneity, Many Labs 2, replication, perspectivism, context

Journal of Social and Political Psychology, 2020, Vol. 8(1), 25–29, https://doi.org/10.5964/jspp.v8i1.1187

Received: 2019-03-18. Accepted: 2019-10-04. Published (VoR): 2020-02-28.

Handling Editor: Debra Gray, University of Winchester, Winchester, United Kingdom

*Corresponding author at: Yale Program on Climate Change Communication, Yale University, 205 Prospect St., New Haven, Connecticut, United States of America. E-mail: matthew.goldberg@yale.edu

Consistency of study results is paramount to scientific research. For example, if the effect of an intervention is considered successful in some studies and not others, the generalizability of the treatment effects remains unclear (Higgins, Thompson, Deeks, & Altman, 2003). This is especially important when considering pervasive failures to replicate the results of original studies (Open Science Collaboration, 2015). In the current commentary, we argue that heterogeneity of effects should be embraced and treated as something to be understood. We explain how a perspectivist approach (McGuire, 1989, 2004) to psychological science that places research questions in an appropriate historical, political, or cultural context is a fruitful path forward for understanding heterogeneity of study results.

In a large-scale replication effort, Klein et al. (2018) test, among other things, whether effects using nearly-identical methods and procedures vary by sample and setting. This large team of researchers aimed to replicate 28 effects published in the field of psychology with each effect tested across approximately 60 samples recruited from 36 different countries or territories. One key conclusion of the project was that “variability in observed effect sizes was attributable more to the effect being studied than to the sample or setting in which it was studied” (p. 446).

The data clearly show that effect sizes vary strongly depending on the topic of study. However, heterogeneity in effect sizes *within* the effect being studied seems to be dismissed as inconsequential. For example, the authors note "heterogeneity across samples does not provide much explanatory power for failures to replicate."

In contrast, here we argue that the variability in effect sizes observed in Many Labs 2 (ML2; Klein et al., 2018) is highly consequential and may in fact explain failures to replicate. Additionally, we describe approaches to the research process that should lead to greater insights from large-scale replication projects such as ML2.

Heterogeneity in Many Labs 2

Although the authors report multiple measures of heterogeneity, here we focus on I^2 for concision and ease of interpretation. I^2 "describes the percentage of total variation across studies that is due to heterogeneity rather than chance" (Higgins et al., 2003, p. 558). This metric runs from 0% to 100%, with 0% indicating no observed heterogeneity and greater values indicating greater heterogeneity.

As the ML2 authors report, 12 of 28 effects (about 43%) exhibited medium or high heterogeneity (see Higgins et al., 2003). Typically, multiple tests of the same phenomenon include different methods and procedures, and therefore heterogeneity should be expected (Higgins et al., 2003). However, considering that great efforts were made to ensure the materials and procedures for all studies in ML2 were identical, it is noteworthy that about 43% of effects still showed substantial heterogeneity. Because materials and procedures were nearly identical, this means that the observed heterogeneity is measuring the lower bound of heterogeneity (McShane, Tackett, Bockenholt, & Gelman, 2018).

A compelling way to investigate the implications of this heterogeneity is to examine the range of observed effect sizes, rather than solely relying on heterogeneity statistics (Borenstein, Higgins, Rothstein, & Hedges, 2015). When viewing Figure 2 in ML2 (p. 470), for example, it is clear that numerous effects—many of which qualified as having successfully replicated—had individual studies whose effects differed in sign. Even for topics where individual studies had entirely or nearly all the same sign, there was still notable variability in effect sizes.

This heterogeneity in effect sizes has clear implications for statistical power. For example, to detect an $r = .1$ effect size with 80% power, a researcher would need to have about 782 participants' data available for analysis. To detect an $r = .5$ effect size with 80% power, a researcher would need only about 28 participants. This effect size range was observed in several cases in ML2. This has large implications for study planning (Kenny & Judd, 2019), especially in a world where resources dedicated to the social sciences are decreasing (Lupia, 2014), and therefore large samples more difficult to obtain.

Additionally, the substantial heterogeneity in effect sizes observed in ML2 has strong practical implications. For example, when social scientific findings are used to make policy decisions, decision-makers need to have a clear understanding of the expected magnitude of an intervention, compare it to alternative interventions, and choose a path forward with limited time and resources. When significant unexplained heterogeneity is present, there is, by definition, less certainty about the consistency of any given effect. Seemingly small differences in effect sizes can translate to large differences in the real world, especially when small events are repeated and can accumulate

over time (Abelson, 1985; Funder & Ozer, 2019). Thus, attempts to explain heterogeneity should be well worth the efforts.

A Path Forward

A fruitful path forward is to use a perspectivist approach to psychological research (McGuire, 1989, 2004). A perspectivist approach "assumes that all hypotheses and theories are true, as all are false, depending on the perspective from which they are viewed, and that the purpose of research is to discover which are the crucial perspectives." (McGuire, 2004, p. 173). Consistent with this perspective, we argue that heterogeneity should be embraced, seen as something to be understood, and used to identify needs for theory development (also see McShane et al., 2018).

When heterogeneity is embraced, researchers can propose variables in advance that might explain effect size heterogeneity, measure them in their study, and aim to explain when effect sizes should be smaller, larger, or run in the opposite direction. For example, Goldberg and colleagues (2019a) recently replicated the same study on a highly contentious issue (climate change) across three sampling platforms (Amazon's Mechanical Turk, Prime Panels, and Facebook), and found significant heterogeneity, with effect sizes for the same manipulation ranging from d = 0.42 to d = 0.86 using a mixed design and from approximately zero to d = .54 using a between-subjects design. The authors explained that significant differences between samples in education, political ideology, and familiarity with the treatment message likely accounted for differences in effect sizes because more educated, liberal, and familiar participants had higher baseline levels of the dependent variable, thereby leading to ceiling effects or sensitization to the treatment (also see Chandler, Mueller, & Paolacci, 2014; Druckman & Leeper, 2012; Goldberg et al., 2019b).

Instead of concluding that between-sample differences are a product of random noise, embracing heterogeneity as something to be understood may lead to fruitful research questions. For example, although Goldberg et al. (2019a) found that demographics and familiarity with the treatment message were promising explanations for effect size heterogeneity, open questions remain as to whether participants from different sampling platforms (e.g., MTurk vs. Facebook) are different in other fundamental ways, or whether different incentives (e.g., paid vs. unpaid participation) can explain differences in effect sizes.

A similar approach can be used to understand heterogeneity in ML2 and other large-scale replication efforts. For example, although the heterogeneity (I^2 = 37%) for Zaval et al. (2014) on heat priming is surprising given the many null effects, an alternative explanation is that the relationship between perceptions of heat and public opinion about climate change is highly context-dependent. For example, Hornsey, Harris, and Fielding (2018) found that the relationship between political ideology and climate change skepticism is emphasized by the political culture in the United States, but this relationship was weak in 24 other countries they examined, thereby pointing to the importance of political culture. A perspectivist approach to the research question would consider whether the findings are limited to certain cultural or political contexts. This aids theory development because it goes beyond the presence or absence of the effect, but rather asks whether the effect exists and in which contexts and under which historical, political, or cultural circumstances the effect is more or less likely to emerge.

In sum, we must remain skeptical of the claim that knowing the effect being studied is necessarily more important than knowledge about the sample and setting. In short, heterogeneity in effect sizes across identical experiments should be used to inform researchers about the boundary conditions of the theories they are testing, as well as the importance of context.

Funding

The authors have no funding to report

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors have no support to report.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*(1), 129-133. https://doi.org/10.1037/0033-2909.97.1.129

Borenstein, M., Higgins, J. P., Rothstein, H. R., & Hedges, L. V. (2015). *I^2 is not an absolute measure of heterogeneity in a meta-analysis.* Unpublished manuscript.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112-130. https://doi.org/10.3758/s13428-013-0365-7

Druckman, J. N., & Leeper, T. J. (2012). Learning more from political communication experiments: Pretreatment and its effects. *American Journal of Political Science, 56*(4), 875-896. https://doi.org/10.1111/j.1540-5907.2012.00582.x

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Goldberg, M. H., van der Linden, S., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. (2019a). The role of anchoring in judgments about expert consensus. *Journal of Applied Social Psychology, 49*(3), 192-200. https://doi.org/10.1111/jasp.12576

Goldberg, M. H., van der Linden, S., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. (2019b). *Convenient but biased? The reliability of convenience samples in research about attitudes toward climate change.* Preprint accessed at https://osf.io/2h7as/

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal, 327*(7414), Article 557. https://doi.org/10.1136/bmj.327.7414.557

Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). Relationships among conspiratorial beliefs, conservatism and climate skepticism across nations. *Nature Climate Change, 8*(7), 614-620. https://doi.org/10.1038/s41558-018-0157-2

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods, 24*(5), 578-589. https://doi.org/10.1037/met0000209

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490. https://doi.org/10.1177/2515245918810225

Lupia, A. (2014). What is the value of social science? Challenges for researchers and government funders. *PS: Political Science & Politics, 47*(1), 1-7. https://doi.org/10.1017/S1049096513001613

McGuire, W. J. (1989). A perspectivist approach to the strategic planning of programmatic scientific research. In B. Gholson, W. R. Shadish, Jr., R. A. Neimeyer, & A. C. Houts (Eds.), *The psychology of science: Contributions to metascience* (pp. 214-245). New York, NY, USA: Cambridge University Press.

McGuire, W. J. (2004). A perspectivist approach to theory construction. *Personality and Social Psychology Review, 8*(2), 173-182. https://doi.org/10.1207/s15327957pspr0802_11

McShane, B. B., Tackett, J. L., Bockenholt, U., & Gelman, A. (2018). Large scale replication projects in contemporary psychological research. *The American Statistician, 73*(1), 99-105.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716.

Zaval, L., Keenan, E. A., Johnson, E. J., & Weber, E. U. (2014). How warm days increase belief in global warming. *Nature Climate Change, 4*(2), 143-147. https://doi.org/10.1038/nclimate2093