Heterogeneity of Effect Sizes: A Response to http://datacolada.org/76 (Heterogeneity is Replicable) by Joe Simmons and Uri Simonsohn

Charles M. Judd and David A. Kenny

Joe Simmons and Uri Simonsohn attribute to us (Kenny & Judd, 2019) a version of effect size heterogeneity that we are not sure we recognize. This is largely because the empirical results that they show seem to us perfectly consistent with the model of heterogeneity that we thought we had proposed. In the following we try to clearly say what our heterogeneity model really is and how Joe and Uri's data seem to us consistent with that model.

Our model posits that an effect size from any given study, d_i, estimates some true effect size, δ_i, and that these true effect sizes have some variation, σ_δ, around their mean, μ_δ. What might be responsible for this variation (i.e., the heterogeneity of true effect sizes)? There are many potential factors, but certainly among such factors are procedural variations of the sort that Joe and Uri include in the studies they report.

In the series of studies Joe and Uri conducted, participants are shown two shapes, one more rounded and one more jagged. Participants are then given two names, one male and one female, and asked which name is more likely to go with which shape. Across studies, different pairs of male and female names are used, but always with the same two shapes.

What Joe and Uri report is that across all studies there is an average effect (with the female name of the pair being seen as more likely for the rounded shape), but that the effect sizes in the individual studies vary considerably depending on which name pair is used in any particular study. For instance, when the name pair consists of Sophia and Jack, the effect is substantially larger than when the name pair consists of Liz and Luca.

Joe and Uri then replicate these studies a second time and show that the variation in the effect sizes across the different name-pairs is quite replicable, yielding a very substantial correlation of the effect sizes between the two replications, computed across the different name-pairs.

We believe that our model of heterogeneity can fully account for these results. The individual name-pairs each have a true effect size associated with them, δ_i, and these vary around their grand mean μ_δ. Different name-pairs produce heterogeneity of effect sizes. Name-pairs constitute a random factor that moderates the effect sizes obtained. It most properly ought to be incorporated into a single analysis of all the obtained data, across all the studies they report, treating it and participants as factors that induce random variation in the effect of interest (Judd, Kenny, & Westfall, 2012; 2017).

Our model of heterogeneity extends things beyond this simple demonstration and would encourage researchers to consider a long list of potential random factors, such as name-pairs, that may moderate effect sizes and induce heterogeneity. For instance, in the studies that Joe and Uri report, name-pairs vary but the two figures used, one with curves and one with jagged edges, always stay the same. By way of replication, one might imagine studies that varied the figure-pairs but kept the name-pair always the same. We expect that similar results would be produced, with different figure-pairs inducing heterogeneity of effect sizes.

Similarly, the studies that Joe and Uri report always recruited participants in the same way, with MTurk appeals. Imagine that other researchers, seeking to replicate the basic effect, used samples recruited in a bunch of other ways. We suspect that different participant samples would have moderated the effect sizes obtained, just like name-pairs were found to and just like figure-pairs might be expected to. Other researchers, convinced all they were doing was attempting to replicate the effect, might use other screen-presentation configurations, with different brightness levels and pixel resolutions, and they too might find that this variation moderated the effect.

The point is that there are a potentially a very large number of random factors that may moderate effect sizes and that may vary from replication attempt to replication attempt. In Joe and Uri's work, these other random factors didn't vary, but that's usually not the case when one decides to replicate someone else's effect. Sample selection methods vary, stimuli vary in subtle ways, lighting varies, external conditions and participant motivation vary, experimenters vary, etc. The full list of potential moderators is long and perhaps ultimately unknowable. And heterogeneity is likely to ensue.

In arguing for heterogeneity (or against it), one can read the available evidence from 'exact' replication efforts in different ways. Joe and Uri read it to say that heterogeneity is minimal. We read it differently. First, we note that the test of heterogeneity (the Q test) is a very low powered test. Reliance on it to detect heterogeneity is relatively insensitive. Second, we admit that when there is no effect to be detected, then heterogeneity ought to be nonexistent. Our reading of the literature is that heterogeneity is obtained in exact replications when the average effect size is other than zero. Finally, it seems to us inappropriate to accept a null hypothesis of no heterogeneity. Surely it is better to grant its possibility and deal with it and its consequences.

As Joe and Uri note in quoting from us, we believe that every replication attempt is necessarily an attempt at generalization. Replications are never exact. They involve slight procedural variations, participant sampling variations, differences in location, climate, participant motivating factors, lighting conditions, experimenters, etc. These all constitute factors that may moderate effect sizes. It is only reasonable to expect that these things matter and, therefore, that heterogeneity is found.

References

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69.

Judd, C.M., Westfall, J., & Kenny, D.A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68,* 601-625.

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, the planning of research, and replication. *Psychological Methods.*