What differences in heterogeneity can tell us about research

T.D. Stanley and Chris Doucouliagos

We are grateful to everyone for their willingness to discuss these important issues and to Andrew for sharing our comments on his website. Joe and Uri's post offers a nice way to address the broader issues that lie at the center of social science research. Last Fall, MAER-Net (Meta-Analysis of Economics Research-Network) had a productive discussion about the replication 'crisis,' and how it could be turned into a credibility revolution. We examined the high heterogeneity revealed by our survey of over 12,000 psychological studies and how it implies that close replication is unlikely (Stanley et al., 2018). Marcel van Assen pointed out that the then recently-released, large-scale, multi-lab replication project, Many Labs 2 (Klein et al., 2018), "hardly show heterogeneity," and Marcel claimed "it is a myth (and mystery) why researchers believe heterogeneity is omnipresent in psychology."

Supporting Marcel's view is the recent post by Joe Simmons and Uri Simonsohn about a series of experiments that are directly replicated a second time using the same research protocols. They find high heterogeneity across versions of the experiment (I^2= 79%), but little heterogeneity across replications of the exact same experiment.

We accept that carefully-conducted, exact replications of psychological experiments can produce reliable findings with little heterogeneity (MAER-Net). However, contrary to Joe and Uri's blog, such modest heterogeneity from exactly replicated experiments is fully consistent with the high heterogeneity that our survey of 200 psychology meta-analyses finds and its implication that "it (remains) unlikely that the typical psychological study can be closely replicated" (Stanley et al., 2018, p.1325). Indeed, we believe that the modest heterogeneity found by ML2 has important implications for understanding research. Because Joe and Uri's blog was not pre-registered and concerns only one idiosyncratic experiment at one lab, we focus instead on ML2's pre-registered, large-scale replication of 28 experiments across 125 sites, addressing the same issue and producing the same general result.

Like previous large-scale replications, only 50% of ML2's experiments are successful (14 of 28), and the replicated effects are, on average, much smaller than the original experiments (.15 vs .60 median SMDs). However, to address critics who blame unwelcomed findings of large-scale replications on the unreliability (or heterogeneity) of their experimental methods and protocols (The Atlantic), ML2 focuses on measuring the "variation in effect magnitudes across samples and settings" (Klein et al., 2018, p. 446). Each ML2 experiment is repeated at many labs using the same methods and protocols established in consultation with the original authors. After such careful and exact replication, ML2 finds only a small amount of heterogeneity remains across labs and settings.[1] It seems that psychological phenomenon and the methods used to study them are sufficiently reliable to produce stable and reproducible findings. Great news for psychology! But this fact does not conflict with our survey of 200 meta-analyses nor its implications about replications (Stanley et al., 2018). In fact, *ML2's findings collaborate both the high heterogeneity our survey finds and its implication that typical studies are unlikely to be closely replicated* by others. Both high and little heterogeneity at the same time? What explains this heterogeneity in heterogeneity?

First, our survey finds that typical heterogeneity in an area of research is 3 times larger than sampling error (I^2= 74%; std dev = .354 SMD). Stanley et al. (2018) shows that this high heterogeneity makes it unlikely that the typical study will be closely replicated (p. 1339), and ML2 confirms our prediction!

Yes, ML2 discovers little heterogeneity among different labs all running the *exact* same replication, but ML2 also finds huge differences between the original and replicated effect sizes (mean absolute difference =.525 SMD). ML2's Table 2 (pp. 468-9) reports 30 original and their replicated effect sizes, measured by Cohen d.[2] The root MSE between original reported effects and the associated replicated effect sizes is a very large (.57 SMD). However, this reflects a total variation that includes sampling errors. ML2's Table 2 reports CIs of both original and replicated experiments, easily converted to sampling variances. We subtract their average sum from the total MSE to get the replication heterogeneity variance (that is, the variance between original and replicated experiments not attributable to random sampling errors in either).[3] Doing so gives the average heterogeneity standard deviation (.4913), which is larger than that reported in Stanley et al. (2018). Wow! This ± .5 SMD deviation is quite large. In fact, it is larger than the typical effect size reported in psychology (Stanley et al., 2018).

If we take the experiments that ML2 selected to replicate as 'typical,'[4] then it is unlikely that this 'typical' experiment can be closely replicated. When this replication heterogeneity has SD=.4913, the probability that original and replicated effects are within .1 SMD is only 16%, 32% for .2, and 46% for ± .3 SMD.[5] With replication heterogeneity as high as this, it is unlikely that any small effect ($.2 \leq d < .5$) or medium-size effect ($.5 \leq d < .8$) will be successfully replicated. Again, confirming our survey.

[1] However, this does not mean that all heterogeneity can be eliminated from psychological experiments. In particular, ML2 found significant heterogeneity in 39% of these 28 experiments, and these tests are known to have low power. ML2 admits that their findings "do not indicate that moderating influences never occur" (Klein et al., 2018, p.484.)

[2] We did not include the two experiments that are reported in units of Cohn's q, because they are not directly comparable. The remaining 26 replicated experiments in Table 2 become 30 because ML2 breaks up four into WEIRD (western, educated, industrialized, rich, and democratic) and less WEIRD subpopulations to control potential cultural heterogeneity, which should reduce heterogeneity.

[3] By subtracting the average sampling variance from the overall MSE, we avoid the problem of truncating at zero and much of the unreliability of that would be found in 30 individual variance estimates.

[4] There is reason to believe these experiments are more replicable than what is actually 'typical,' because ML2 eliminated from consideration those experiments known to be difficult to replicate (Klein et al., 2018, n.1; p.486).

[5] These calculations assume large samples with negligible sampling error; otherwise, these prob's are smaller still.

What explains the large difference between the heterogeneity found among exactly duplicated ML2 replications and the heterogeneity ML2 finds between the original and replicated effect sizes? As suggested by ML2, these differences may be "due to errors in replication design, p-hacking in original studies, or publication bias" (p. 477). Thus, the high heterogeneity, routinely seen in reported research, is largely due to researcher-controllable choices. This has huge implications for turning the current 'crisis' into a credibility revolution. Because ML2 attempted to exactly reproduce the original experiments, the remaining differences must come from what ML2 did not control: publication bias, QRP and unobservable differences in research methods and protocols.

Nor are ML2's findings unique. Virtually all large-scale replications find large differences between the original effects sizes and their replicated effects, which is heterogeneity of a similar scale and type that is found in most meta-analyses and reflected by our survey (Stanley et al., 2018).

Heterogeneity may not be omnipresent, but it is frequently: seen among published research results, identified in meta-analyses, and confirmed by large-scale replications. As Blakeley, Ulf and Karsten reminds us, heterogeneity has important theoretical implications, and it can also be identified and explained by meta-regression analysis.

References:

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 sample and setting. *Advances in Methods and Practices in Psychological Science* 1(4) 443- 490.

Stanley, T.D., Cater, E. and Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*. **144:**1325-46.

Stanley, T.D. and Doucouliagos, C. (2018). Towards a Credibility Revolution: Why successful replication remains unlikely, posted Oct 28, at MAER-Net.org.