Response to:
Datacolada Post [76] "Heterogeneity Is Replicable: Evidence From Maluma, MTurk, and Many Labs"

Blakeley B. McShane, Ulf Böckenholt, and Karsten T. Hansen

The comments we sent to Joe and Uri on April 20, 2019 before their blogpost with the title above went live are available ***below*** (they denied our request to link to this from their blogpost). Our "Large-Scale Replication" paper discusses many of these issues in greater depth (especially on page 101) and is available here. Our comments on Andrew Gelman's blog discusses many of these issues in greater brevity and is available here.

In reading their blogpost, we were struck by the fact that we have already had many in depth discussions with Uri touching on these and other issues in December 2017 – January 2018. This discussion was to culminate in a dialogue on Andrew Gelman's blog which unfortunately did not work out. Instead, Andrew proceeded with his own post and solicited brief comments from each of us. Our longer comments from that conversation are available here (see especially the Empirical and Redux sections) and Andrew's blogpost which includes our brief comments (see especially our point [5]) are available here.

April 20, 2019

Response to:
Datacolada Post [76] "Heterogeneity Is Replicable: Evidence From Maluma, MTurk, and Many Labs"

Blakeley B. McShane, Ulf Böckenholt, and Karsten T. Hansen

We thank Joe and Uri for featuring our papers "Adjusting for Publication Bias in Meta-analysis" (APB) and "Large-Scale Replication in Contemporary Psychological Research" (LSR) in their recent blogpost. We write to clarify some perceived misrepresentations of our views reflected in the blogpost and to show that a strong prior belief in homogeneity undergirds the analyses and interpretation thereof presented in the blogpost. The blogpost has two main parts, one on the new Maluma-Takiti data of Joe and Uri and one on prior "many labs" data. As our published work concerns the latter, we begin by discussing that.

Before proceeding, however, we wish to clarify one major issue. While our work has, in line with a long tradition, argued that effect size heterogeneity is unavoidable in psychological research, we have never said this means "that we should not expect individual studies to replicate" (or even that heterogeneity is problematic). It is obvious and uncontroversial that heterogeneity impacts replicability (no matter how it is operationalized in terms of study design, statistical findings, etc.), and thus we have merely advocated that it is one of the many things that must be accounted for in study design and statistical analysis—whether for replication or more broadly.

"Many Labs" Studies
[Note: Much of this section is given in greater detail on page 101 of LSR.]

We previously reported average heterogeneity of $I^2 \approx 40\%$ (or $\tau \approx 0.2$) in the Many Labs studies as whole (see Table 3 of the Many Labs paper) and $I^2 \approx 20\%$ (or $\tau \approx 0.1$) in the M-Turk subsample of these studies (our calculations). Substantively, we view this nonzero heterogeneity as notable because these studies were explicitly designed to eliminate heterogeneity. Yet, $\tau = 0.2$ means it would not be surprising for the true size of some effect to be, say, $\delta = -0.1$ in one lab and $\delta = 0.3$ in another.

We also view it notable for purely statistical reasons. Specifically, in data settings like this, the statistical models (i.e., estimators and tests) used to arrive at these figures have poor statistical performance properties that tend to falsely suggest zero heterogeneity (e.g., low power, implausibly frequent estimates on the boundary of zero; see page 101 for details). Thus, only one with a strong prior belief in zero heterogeneity would take such results favoring zero as confirmatory of that belief. Indeed, we have described (multilevel multivariate) statistical models with superior performances properties and suggested they be used for this and similar data when, as in the blogpost, it is unnecessary to exactly mimic the Many Labs authors' statistical model as it was in our published work (again, see page 101).

In sum, we have always viewed these figures cautiously due to the statistical models upon which they are based but find the nonzero results interesting (even surprising) for both substantive and statistical reasons.

Such caution should obviously be extended to any results based on these models—including the blogpost simulation. Nonetheless, even bearing this caution in mind, the blogpost simulations impose a strong prior belief in zero heterogeneity. Specifically, they posit (indeed privilege) a belief in zero heterogeneity and only reject it given sufficient evidence to the contrary. However, because heterogeneity has long been considered the norm in psychological research, this approach seems backwards, and it seems much more reasonable to do the opposite: assume some degree of heterogeneity and only conclude it is zero given sufficient evidence for that. At minimum, one should at least consider a range of values—not zero alone.

Also, the blogpost finding that estimates of I^2 (or τ for that matter) are biased upward when heterogeneity is zero is practically tautological: estimates of I^2 (and τ) are necessarily greater than or equal to zero and

thus can only be biased upwards when the truth is zero as in the blogpost simulation. More interesting to someone who does not have a strong prior belief in zero heterogeneity is the fact that estimates of I^2 are biased downwards—to a considerable degree—even for true I^2 as low as 0.3 for data settings with few studies as in the blogpost simulation (von Hippel 2015; doi:10.1186/s12874-015-0024-z). This could be used to support the view that true heterogeneity may well be higher than the 20% estimated—exactly opposite the conclusion of the blogpost!

Finally, our single sentence mentioning Eerland et al. and Hagger et al. was meant as illustrative not as a comprehensive review (presumably like Table 1 of the blogpost). However, we again urge caution in interpreting these results: all of these papers employ the estimators and tests with known poor performance properties discussed above. Thus, results must be regarded with caution, particularly given the large number of boundary estimates of zero reported in the table which create problems for statistical inference. As a small correction, we note the two p-values of one reported in the blogpost are not reflected in Eerland et al. (we believe these are 0.662 and 0.393 respectively).

Maluma-Takiti

While the Maluma-Takiti findings are obviously just one illustration from one paradigm, they provide much fodder for thought. We provide three.

First, the fifty wave one studies depicted in Figure 1 of the blogpost are obviously not independent as each of the 508 subjects participated in ten studies each; the same holds for the fifty wave two studies. Perhaps surprisingly, however, it may also be the case that the two studies in each wave one / wave two pair depicted in Figure 2 of the blogpost are not independent depending on the blocking employed. Consequently, the scatter about the line on view in that figure may understate the true degree of variation.

Second, while someone with a strong prior belief in homogeneity might look at the scatter on view in the figure and see consistency, we see substantial scatter (despite any potential understatement discussed above), with many pairs of d's differing by 0.3 or even 0.4! So, the figure arguably fails to support the claims in the blogpost. In any event, a proper multilevel multivariate analysis that accounts for the various forms of dependence among the individual-level observations is necessary to ascertain such a thing and so we humbly request that Joe and Uri make this data public so it can be subject to such an analysis.

Third, we comment on two aspects of how Joe and Uri operationalized replication between the waves:

(i) *Study Design*: One choice Joe and Uri appear to have made was to use the same fifty pairs of names in the two waves. This seems sensible, but of course another reasonable choice would have been to have used the top fifty names for each wave rather than repeat the top from the first. Since the waves were only two years apart and top names are probably pretty stable from year to year, it likely does not matter much for this example. However, we would love to hear Joe and Uri's thoughts on why they chose this operationalization, how they might have thought otherwise were the waves more separated in time, and what general template they might offer researchers seeking to conduct replication studies outside this paradigm. We would appreciate this practical guidance.

(ii) *Statistical Findings*: We were delighted to see Joe and Uri eschew the typical operationalization of replication (i.e., whether or not both studies matched in terms of sign and statistical significance) and instead focus directly on the pairs of effect size estimates. While a more direct measure than the correlation coefficient would have been our choice, this is a big step forward for them and the field!