Replication and Effect Size in Psychological Science: Comments Inspired by Gilbert et al. (2016)

D. Stephen Lindsay

The Open Science Collective (OSC; 270 authors led by Brian Nosek) reported attempts to replicate 100 psychology experiments in a 2015 article published in Science. In March 2016 a Technical Comment on that project by Gilbert, King, Pettigrew, and Wilson appeared in Science. They concluded that “OSC’s data clearly provide no evidence for a ‘replication crisis’ in psychological science.” A subset of the OSC authors coauthored a reply to Gilbert et al. in Science, and subsequently Brian Nosek and Elizabeth Gilbert (no relation to Daniel Gilbert, lead author of the Technical Comment) posted a piece on Retraction Watch. Several other responses to the Gilbert et al. Technical Comment have been posted on blogs (e.g., Sanjay Srivastava, Uri Simonsohn, Daniel Lakens, Simine Vazire, Andrew Gelman, David Funder, Rolf Zwaan, and Dorothy Bishop). Gilbert et al. posted responses to some of the criticisms.

Given that so many pixels have already been spilled on this matter, it may be silly of me to think I have anything to add. Certainly I will leave assessment of Gilbert et al.’s statistical approach to experts. But I would like to share a few more general thoughts inspired by the Gilbert et al. Technical Comment.

First, I hope that many psychology researchers will read the OSC (2015) article as well as the Gilbert et al. (2016) Technical Comment and at least some of the online discussions of those articles. These are important, timely, and multifaceted issues that demand thoughtful attention.

Gilbert et al. (2016) implied that OSC (2015) is the wellspring of concerns about a replication crisis. But such concerns have been widespread for several years and for many reasons. The OSC article brought renewed attention to those concerns but it is not their source.

Gilbert et al. (2016) repeatedly referred to “the” benchmark OSC used to adjudicate the success or failure of replication attempts. The OSC article did not use a single benchmark to assess which replication attempts “succeeded.” To my eye, their article was sophisticated and nuanced in its treatment of the complex issue of assessing whether or not a replication succeeded. But some media reports simplified that message and indeed got it fundamentally wrong by saying that the results showed that most of the effects tested by OSC “could not be” replicated. Such a statement insinuates that the effects that did not meet alpha .05 in the replication attempts were bogus Type I errors in the original publications. That is not at all what the OSC article said. It is impossible for a replication attempt to show that an effect “cannot be” replicated, because it is always possible that under other conditions (e.g., with larger samples, less noisy measures, stronger manipulations, tighter controls, and/or better luck) the effect would be replicated.

From my perspective, the best summary of the OSC findings is that most replication attempts yielded estimates of effect size that were smaller than those in the original studies. OSC found
that on average the effects in replication attempts were about half the size of those found in the original articles. That observation converges with a number of other reasons to believe that common practices among researchers, reviewers, and journal editors tend to inflate estimates of effect size. A simple example is publication bias, the practice of reporting only the studies that yielded the largest effect, leaving other studies in the file drawer. It is likely that for many of the effects that failed to replicate in the OSC project there really is an effect but that effect is much smaller than the effect reported in the original article. OSC projects set $N$ to achieve 90% power, but those calculations may often have been based on inflated estimates of effect size, such that power to detect the real (smaller) effect was much lower than 90%. That is why high-powered replications such as the Many Labs Project more often obtained evidence of effects (if only in analyses combining findings across studies). Gilbert et al. might protest that this is a central part of their argument, but my point is that it would not be news to close readers of OSC (2015).

The OSC replication attempts differed from the original studies in numerous ways (e.g., different labs, experimenters, subjects, years). Effects that are greatly attenuated by such differences are by definition of limited generalizability. The Gilbert et al. Technical Comment highlighted the fact that some of the OSC replication attempts differed from the original studies in substantive ways that would reasonably be expected (on theoretical grounds) to reduce or eliminate the effect. The OSC article acknowledged variability in the fidelity of replications and attempted to assess the extent to which such differences contributed to differences between the results of original versus replication studies (e.g., by asking original authors to assess the fidelity of the replication). This will probably be a live issue for some time as scientists attempt to unravel the contributions of substantive differences versus statistical shortcomings in producing differences in effect-size estimates between original studies and replication attempts.

The first Registered Replication Report (Alogna et al., 2014) starkly illustrates some of the issues discussed above. This multi-lab direct replication was inspired by Schooler and Engstler-Schooler’s (1990) research on “verbal overshadowing.” In S&E’s Experiment 1, subjects watched a video of a bank robber and later attempted to identify the robber in a lineup. Between the video and the lineup, some subjects did nothing but filler tasks, whereas other subjects were required to write a detailed description of the robber’s face (as well as to perform some brief filler tasks such that delay between video and test was equated across conditions). The robber was correctly identified by 64% of the control subjects and only 38% of face-verbalization subjects, a statistically significant effect replicated in other experiments reported by S&E and sometimes but not always in subsequent follow-up studies. The RRR had intended to replicate S&E’s Experiment 1, in which subjects in the verbalization condition watched the video, performed filler tasks, described the robber’s face, and then took the test immediately after verbalization (VFVT), but initially they instead had subjects verbalize immediately after the video and then perform the filler tasks before taking the test (VVFT). With the VVFT order, zero of 31 labs obtained a statistically significant verbal overshadowing effect. You might think that 31 failures to replicate would be crushingly strong evidence for the null hypothesis. But most of the experiments obtained directionally better performance in the
control than verbalization condition, and a meta-analysis across the 31 lab replications indicated that there was a small (4%) but statistically significant effect. Moreover, 22 of the labs subsequently repeated the experiment with the VFVT order. Even with that order, only 8 of those studies yielded a statistically significant difference, but the tendency was in the predicted direction in every study and the meta-analysis estimate of effect size was 16%. That is quite a bit smaller than the effect reported by S&S-E, but it is not trivial and the consistency across the 22 replications is quite compelling. These results justify great confidence that Verbal Overshadowing is (at least with the VFVT order) a real, replicable albeit small effect, despite the fact that small-N tests of that effect often fail to reject the null hypothesis.

The Gilbert et al. Technical Comment may help correct the misapprehension that the OSC results indicated that most effects published in psychology are bogus. That would be good. And it may help to promote efforts to estimate the contributions of statistical shortcomings versus hidden moderators to differences in the effect sizes in original studies versus replications. That too would be good.

I worry, though, that the Gilbert et al. article may have the unfortunate and unintended side effect of leading people to believe that there are no problems with “business as usual” and that it is appropriate to conduct lots of studies with noisy measures and smallish samples (and with great latitude as to how data are analyzed and interpreted after they are obtained) and to then report the subset of studies and analyses that yielded provocative results. There is broad consensus among statisticians/methodologists that those practices should be discouraged. The Gilbert et al. Technical Comment does not lessen my belief in the importance and value of such reforms.