More on “Estimating the Reproducibility of Psychological Science'”[1]

Daniel Gilbert, Gary King, Stephen Pettigrew, and Timothy Wilson

7 March 2016

Summary

Our Technical Comment on OSC2015, the OSC’s reply to it, and our reply to the OSC, have elicited lengthy responses from several colleagues. We were grateful to hear their thoughts and to engage them in this important conversation. For those who have not been following the conversation, here is how it has gone so far:

> OSC: “We have provided a credible estimate of the reproducibility of psychological science.”
>
> US: “No, you haven’t, because (1) you violated the basic rules of sampling when you selected studies to replicate, (2) you did unfaithful replications of many of the studies you selected, and (3) you made statistical errors.”
>
> OSC (& OTHERS): “We didn’t make statistical errors.”

We still think they made statistical errors and we explain why below, but before we do it is important to note that they while some colleagues have challenged our Point 3, none has challenged our Points 1 or 2, probably because it requires no special expertise to see that these points are inarguable facts. And yet, these inarguable facts are *by themselves* sufficient to repudiate the OSC’s claim. We believe we can convince readers that Point 3 is also an inarguable fact, but whether or not we succeed, our conclusion that OSC2015 does not provide a credible estimate of the reproducibility of psychological science is inescapable—and remains the one and only claim of our Technical Comment. All the other claims with which critics may wish to argue—that psychological science is 100% replicable, that publication bias does not exist, that replications are unimportant, that data sharing rules cannot be helpful, and so on—are claims we never made and do not believe.

We now explain Point 3 by addressing the comments of several critics. For those who do not wish to wade through the details of our statistical arguments, our bottom line is this: Although many (but not all) of these critics’ claims turn out to be incorrect or immaterial, granting them would simply *strengthen* our conclusion.

[1] Copies of our work on this subject can be found at http://j.mp/PsychRep.

Details

First, in our Technical Comment, *we made one and only one key claim.* Although journalists, scientists, policy-makers, and lay people worldwide concluded that OSC2015 showed that the replicability of psychological science is low, the evidence provided in OSC2015 does not support such a conclusion. We do not claim that replicability is high, and it is clearly not true that "Gilbert et al claim that original studies would replicate just fine if only replicators would get the procedures right" (see blog post by Sanjay Srivastava). Our only claim is that, on the basis of the evidence collected by OSC2015, one cannot distinguish whether the replicability of psychological science is high or low. The uncertainty is much larger than indicated in OSC2015. Another study with a better research design may well be able to do so, but this one could not. The remainder of this document, like our other work on this topic, addresses only this point. We understand that some people believe that some of the surprising results in psychology are theoretical nonsense, knife-edged, p-hacked, ungeneralizable, subject to publication bias, and otherwise unlikely to be replicable or true (e.g., see blog posts 1 and 2 by Andrew Gelman), but in our Technical Comment we took no positions on these issues, and we will not do so in this document.

Second, the target quantity of interest is the replicability of psychological science, of which several measures exist (CIs, hypothesis tests, p-values, etc, etc.). Some of our critics fault us for discussing only one of these measures. But we chose to discuss only one of these measures for the simple reason that *the same conclusions result no matter which (reasonable) measure of replicability one chooses*. The different measurements are based on different theories of inference, but there is no significant disagreement among them with regard to any analyses we did or that the authors of OSC2015 did. We chose to discuss the measure that made the most sense (as described in our next point) and that is most easily communicated, but that choice is immaterial to our key claim. Our conclusions are also unchanged if we "reverse time" and consider whether the original studies replicate those conducted by OSC2015.

Third, *a scientific statement is not one that is necessarily true; it is one that is made with an honest estimate of the degree of uncertainty*. The less uncertainty, the bigger the claim, and the more weight an author is claiming we all should put on his or her conclusions. For example, a study with only 3 observations (but measured well, drawn from a known population, with accurate uncertainty estimates, etc.) makes a small contribution, but a contribution nonetheless. Thus, when evaluating a study, one must condition on the level of uncertainty of the original study's inference. We do this here for OSC2015, and any individual replicator must do this for any result they attempt to replicate. This is why it usually makes the most sense to evaluate whether a point estimate from the replication falls within the claimed CI of the original study. Relatedly, Srivastava points out: "Another issue that is critical to interpreting intervals is knowing that intervals get wider the less data you have. This is never addressed, but the way Gilbert et al. use original studies' confidence intervals to gauge replicability means that the lower an original study's power, the easier it will be to 'successfully' replicate it." This is indeed true, but it is a feature, not a bug. Studies do not need to be weighted by how much uncertainty they come with when replicating their results because the point of a replication exercise is to judge the accuracy of the uncertainty put forward by the author of the original study. In that exercise, the original author's given level of accuracy is fixed. Similarly, "publication bias" (aka, the "file drawer problem") may well affect some or even many published studies, but it is not relevant to OSC2015 or to our comment on it, and it does not affect calculations in either because the analysis in each is about the published literature and thus conditions on the scientific statements made in the original studies.

Most importantly, this argument about sample size further supports our conclusion. OSC2015 used a 95% (or 92%) baseline of the percent of studies that would replicate due to (sampling variability-based) chance alone. When using data from the Many Labs study to estimate the correct figure, the results indicate a success rate due to chance alone of 34.0%. When we rerun the analysis, separating the replications based on whether they have a larger or smaller sample than the original study, the success rate varies with sample size in the expected direction. Of the Many Labs studies that had a smaller sample size than the original studies that they replicated, 25.2% of replications fell inside the original study's CI. Of those that had a sample size equal to or larger than the original study, the expected success rate was 44.7%. Yet, each of these estimates is even lower than OSC2015's observed success rate replicating original studies (47%), thus indicating insufficient evidence to conclude that OSC2015's replication success rate is any lower than should be expected by chance alone.

Fourth, *to have any scientific meaning, point estimates of population parameters must come with a baseline standard of comparison or "benchmark."* For example, OSC successfully replicated 47% of the original studies (based on CIs). The benchmark here is the percent of replication successes we would expect due to chance alone. So the question is how this benchmark is computed. In one place, OSC2015 assumes that the benchmark should be 8% (the average of their power calculations, which assumes their replications differ only due to sampling variability); for other metrics and purposes, OSC2015 fails to report any benchmark at all. Now, obviously, (1) a "confidence level" refers to the percent of samples from the same population that capture the truth, and so (2) the degree to which an imperfectly estimated CI from one sample captures a point estimate from another sample from the same population will be *lower* than 95% (Srivastava). Our Technical Comment begins with OSC2015's chosen high benchmark (either 92% or 95%) and shows that a correct calculation of this benchmark—that is, one that takes into account the sources of variability that they ignored—is substantially lower. Some have argued that 95% (or 92%) is too high a benchmark for the CI metric because of (2) above. For example, Lakens suggests 83.4% as a benchmark while still assuming only sampling variability. We have no problem with this argument, *but accepting it means that OSC underestimated the replicability of psychological science even more drastically than the calculations in our Technical Comment indicated.* Of course, the correct benchmark must recognize all genuine sources of variability and cannot be limited to the fiction that the only way in which the replications differed from the originals was that they drew different subjects from the identical population. We showed that each deviation from the experimental protocol of the original studies upwardly biased their estimate of the expected success rate. When we included some of these factors in a more appropriate calculation, the benchmark became lower, thus supporting our key claim.

Fifth, some phenomena are not very robust and occur only when very precise conditions are met. One example is human life, which is so fragile that it exists only in a very narrow range of temperature, oxygen concentration, atmospheric pressure, and so on. But questions about robustness are different than questions about replicability—which is about whether a specific phenomenon re-occurs when the specific conditions described in the original study are faithfully reproduced. *Robustness to changes in conditions* *is important, but it is not relevant to the replicability of psychological science which is what OSC2015 tried to estimate.* The claims we evaluate in our Technical Comment are those in OSC2015, not those in the original studies. As we explain, OSC2015 is a meta-scientific exercise in which the unit of analysis is a study rather than a human research subject, and therefore it must obey the rules of science. To evaluate OSC2015's claim, we thus need to estimate the variance (or CI) of quantities such as the percent of studies that would replicate by chance alone or the percent that did replicate successfully using OSC2015's procedures, across hypothetical repeats of the entire OSC2015 study. This variability—by definition—includes sampling variability as well as every other step OSC took. These steps include the following features hard coded into the design of OSC2015: Replicators were given guidance but were also given many degrees of freedom to select the articles, and studies within the articles, they wished to replicate. They were directed to choose the last study in each article, but they deviated from this rule in some instances. They were also allowed to change both the experimental methods and the populations sampled as they wished. Although they consulted the original authors about these changes, not all authors agreed with them. Because the replicators could each make so many independent decisions, we must also include as a source of variability in this process the choices made by the replicators who got to determine the answers. All of these factors increase the variance well beyond that computed in OSC2015 based solely on sampling variability. We estimated the actual variance by using data from the Many Labs study (as suggested to us by OSC2015's corresponding author, Brian Nosek). It is worth noting that our estimates are almost certainly *under*estimates because the Many Labs study locked down experimental protocols with much more precision than did OSC2015, but at least it gives us an empirical estimate. As Simonsohn writes, "For readers to consider whether design differences matter, they first need to know those differences exist. I, for one, was unaware of them before reading Gilbert et al." Srivastava agrees: "Some of the protocol differences between originals and replications deserve closer scrutiny, and it is good that Gilbert et al. brought them to our attention."

Sixth, in our Technical Comment, we check whether the replication point estimate is inside the CI of the original study (Criterion 1). Others have suggested that a more appropriate criterion is whether the original point estimate is inside the CI of the replication (Criterion 2). While Criterion 1 is usually the correct one (see the Third point above), it barely matters because our conclusions are identical in both cases. When we use the Many Labs replications to estimate the number of expected replications that succeeded—by iteratively treating each Many Labs "replication" as a "published" result and comparing the remaining 35 replications to it—we arrive at precisely the same estimate, 65.5% successes, using either criterion. This makes sense given that, for this calculation, the "published" and "replication" results are being drawn from the same pool of studies. In a separate calculation, we compare the results from the Many Labs data to its original studies. We find a slight difference when applying the two criteria, but with no change in conclusions: In our Technical Comment, we report, using Criterion 1, that the observed success rate in Many Labs is 34.0%. When we apply Criterion 2, the expected success rate is 39.4%.

Seventh, although different measures of replicability lead to the same conclusions, Srivastava is right when he says that one of our figures can be interpreted as making an inappropriate cross-measure comparison. Using our CI metric alone, 58.3% of the original studies in Many Labs successfully replicated when pooling labs (or 40% when pooling subjects, a lower variance but less robust approach). But, as we reported, only 34.0% successfully replicated when the data in each lab is analyzed independently. For the “sign and significance” measure, 84.6% of studies replicate when the data are pooled, but only 67.6% replicate when analyzed individually. We appreciate Srivastava’s correction and note the reduction in the effect size, but our conclusion that OSC2015 underestimated the reproducibility of psychological science still holds (even ignoring all other problems we raised, any one of which was also sufficient to lead to this conclusion).

Eighth, a few critics have discussed the implications of how OSC2015 selected the articles and studies to replicate. We did not have room to discuss this issue in our Technical Comment, but we did include some discussion of it in pages 7-9 of our response to OSC’s reply to our Technical Comment. Details appear there, but the essential point is that in order to use the sample analyzed in OSC2015 to infer the population of research studies in psychological science, we need a probability sample with known characteristics. Unfortunately, OSC2015 did not have anything even remotely approximating a probability sample. Their sample was self-selected with few fixed or known characteristics. Even their *intended* sample—the last study in every article published in 2008 in three journals—is not representative of all of psychological science. Moreover, the replicators were allowed to deviate from the rules they set for themselves, were allowed to exclude any study that might require more time or money than they had, and so on. Quite apart from the infidelities in the reproduction of methods, this fact alone is enough to keep OSC2015 from providing a credible estimate of of the reproducibility of psychological science.

At the end of the day, OSC2015 represents a massive organizational effort that provides a good deal of information about the robustness of each of the 100 individual studies they examined. However, the collective evidence offered in OSC2015 does not support the conclusions about embarrassingly low degrees of replicability in psychological science that many have drawn from it—and indeed, it does not even support the OSC2015’s claim to have provided a credible estimate of the reproducibility of psychological science. Srivastava notes that “The RPP is not perfect...but all science proceeds on fallible evidence (there isn’t any other kind).” We agree there isn’t any other kind. But just because all evidence is fallible does not mean that all evidence is equally fallible. Srivastava claims that OSC2015 “gave us a sense, however rough, of where the field stood.” An estimate this rough is no estimate at all, and indeed, neither we nor our critics would turn an equally blind eye to the major flaws of any other study. Although we admire the intentions of OSC2015, its conclusions are unwarranted.