A RESPONSE TO THE REPLY TO OUR TECHNICAL COMMENT ON "ESTIMATING THE REPRODUCIBILITY OF PSYCHOLOGICAL SCIENCE"

Daniel T. Gilbert[1], Gary King[1], Stephen Pettigrew[1], Timothy D. Wilson[2]
Harvard University[1], University of Virginia[2]

We thank our friend Brian Nosek and his 43 colleagues for their thoughtful reply (hereafter referred to as OSC-Reply) to our Technical Comment on their article "Estimating the reproducibility of psychological science" which recently appeared in *Science* (hereafter referred to as OSC-2015).

Although some differences between us remain, we are delighted that the authors of OSC-Reply agree that OSC-2015 provides no grounds for drawing "pessimistic conclusions about reproducibility" in psychological science. We hope they will work as hard to correct the widespread public misperceptions of their article as they did on the article itself.

In our Technical Comment, we demonstrated that the methodology and statistical analyses of OSC-2015 were flawed, and that when those analyses are corrected for *error*, *power*, and *bias*, the pessimistic conclusions that many have drawn from this article—namely, that there is a "replication crisis" in psychology—are unwarranted. Indeed, as the authors of OSC-Reply acknowledge, the evidence in OSC-2015 is perfectly consistent with the opposite conclusion.

The authors of OSC-Reply did object to some of our arguments, which we review in detail below. Perhaps our biggest disagreement is about how closely the replication studies followed the procedures of the original studies. The authors of OSC-Reply correctly note that there is no such thing as an "exact replication" because there are always *some* differences between the conditions under which an original and a replication study are performed. As they noted, "They are conducted in different facilities, in different weather, with different experimenters, with different computers and displays, in different languages, at different points in history, and so on." This is true. But just because original and replication studies are not identical does not mean that the extent of their differences is irrelevant. And indeed, anyone who spends some time reading the replication reports that were the basis of OSC-2015 will see that many of the replication studies were very different from the originals. These differences are not the trivial ones that the authors of OSC-Reply mention, such as different buildings and different weather. These are differences that the studies themselves (as well as a long history of psychological research) indicate should and did matter. We offered a few examples in our Technical Comment and we provide another striking example below, but the interested reader will have no problem finding more.

Why does the fidelity of a replication study matter? Even a layperson understands that the less fidelity a replication study has, the less it can teach us about the original study. That barely needs to be said. But there is another reason why fidelity matters and it *does* need to be said, which is why we devoted a large part of our Technical Comment to saying it: The fidelity of a set of replication studies changes the number of replication studies that one should expect to fail by chance alone. If one does not properly estimate how many replication studies should fail by chance alone, then the number that actually do fail is utterly uninformative. Infidelity is a

measurable source of error, and the authors of OSC-2015 did not measure this source of error and did not even try to take it into account, which means that they could not have correctly estimated how many of their replication studies should have failed by chance alone.

In sum, the authors of OSC-2015 allowed many of the replication studies to depart significantly from the methods and samples of the original studies, and then they included the results of these low fidelity studies in their analyses even though there was clear evidence that doing so lowered their estimates of reproducibility. Because of these errors and omissions, numerous psychologists, journal editors, policy makers, and journalists have mistakenly interpreted the OSC-2015 results to mean that psychological science is deeply flawed.

We turn now to a detailed discussion of OSC-Reply's objections to our arguments. We conclude with a discussion of another important limitation of the OSC-2015 design that we did not have the space to address in our Technical Comment.

OUR RESPONSE TO OSC-REPLY

OSC-Reply begins with a general criticism of our focus on confidence intervals as a measure of reproducibility:

> *GKPW focused on a variation of one of OSC-2015's five measures of reproducibility - how often the confidence interval (CI) of the original study contains the effect size estimate of the replication study.*

This is a red herring. Neither we nor the authors of OSC-2015 found any substantive differences in the conclusions drawn from the confidence interval measure versus the other measures. We focused on CI for simplicity and ease of interpretation.

We organize OSC-Reply's remaining objections around the three arguments in our Technical Comment: *error*, *power*, and *bias*.

Error

In order to determine how many of the original studies reported true findings, it is necessary to estimate how many replications should fail due to sampling error alone. The authors of OSC-Reply questioned our statement that "sampling error alone should cause 5% of the replication studies to 'fail' by producing results that fall outside the 95% confidence interval." They wrote:

> *GKPW misstated that the expected replication rate assuming only sampling error is 95%, which is true only if both studies estimate the same population effect size and the replication has infinite sample size (2,3). OSC-2015 replications did not have infinite sample size. In fact, the expected replication rate was 78.5% using OSC-2015's CI measure (see OSC-2015's SI p. 56, 76, https://osf.io/k9rnd/). By this measure, the actual*

replication rate was only 47.4%, suggesting the influence of factors other than sampling error alone.

There are different ways to calculate the amount of error that should be expected when replicators draw a new sample from the same population and follow the same procedures as the original study, and the authors of OSC-Reply suggest an alternative way of estimating that error. But two things should be noted. First, their method results in a *larger* error rate than ours does; that is, their method *leads to an increase* in the number of replication studies that should be expected to fail by chance alone and therefore strengthens our claim. Second, the authors of OSC-Reply fail to address one of the central arguments in our Technical Comment, namely, that replicators *did not simply draw new samples from the same populations and follow the original procedures.* Indeed, in many cases replicators used remarkably different populations and remarkably different procedures, thereby introducing a new source of error that the authors of OSC-2015 did not even try to estimate.

So we did. In order to estimate this additional source of error, we analyzed data from one of the "Many Labs" studies sponsored by the OSC. Brian Nosek was kind enough to suggest this to us and referred us to these projects. Our analysis indicated that the actual replication rate in the OSC-2015 data was considerably higher than the authors of OSC-2015 recognized. The authors of OSC-Reply raised several objections to our analysis:

Within another large replication study, "Many Labs" (4, ML2014), GKPW found that 65.5% of ML2014 studies would be within the confidence intervals of other ML2014 studies of the same phenomenon and concluded that this reflects the maximum reproducibility rate for OSC-2015. Their analysis using ML2014 is misleading and does not apply to estimating reproducibility with OSC-2015's data for a number of reasons.

First, GKPW's estimates are based on pairwise comparisons between all of the replications within ML2014. As such, for roughly half of GKPW's failures to replicate, "replications" had larger effect sizes than "original studies" whereas just 5% of OSC-2015 replications had replication CI's exceeding the original study effect sizes.

This point would be probative only if the replication studies had faithfully followed the procedures of the original studies and differed *only* in the samples they drew from the same population. As we explained in our Technical Comment (and as we expand upon below), many of the replication studies used *very* different populations and *very* different procedures than the original. This could explain why the effect sizes of the replication studies were smaller on average than the effect sizes of the original studies.

Second, GKPW apply the by-site variability in ML2014 to OSC-2015's findings, thereby arriving at higher estimates of reproducibility. However, ML2014's primary finding was that by-site variability was highest for the largest (replicable) effects, and lowest for the smallest (non- replicable) effects.

This may simply be a trivial statistical artifact. The variance of large numbers is almost always larger than the variance of small numbers. But regardless, the existence of such a correlation is not material to the claims in our Technical Comment.

> *If ML2014's primary finding is generalizable, then GKPW's analysis may leverage by-site variability in ML2014's larger effects to exaggerate the impact of by-site variability on OSC-2015's non-reproduced smaller effects, thus overestimating reproducibility.*

We used the ML2014 data to ask the question, "If an original study reported a true effect, and if many different laboratories tried to reproduce that effect, what kind of variation in findings would occur by chance alone?" The above comment by the authors of OSC-Reply actually reinforces our point that such variation would be large, and that this variation was not taken into account by the authors of OSC-2015.

In summary, the authors of OSC-Reply quibble with us about how best to correct for a potent source of error, but they overlook the main point, which is that the authors of OSC-2015 used *no* method to correct for this potent source of error.

Power

We showed that by attempting to replicate each original study only once, OSC-2015 used a relatively low powered design. In contrast, the Many Labs project used a much higher powered design that involved replicating a handful of original studies 35-36 times each, and they found a much higher replication rate (85%). So we asked what would have happened if the Many Labs project had used the same low-powered design used by the authors of OSC-2015 (i.e., one replication per study), and we found that under these circumstances, the replication rate would have dropped from 85% to 34%. The authors of OSC-Reply raised the following objection:

> *GKPW use ML2014's 85% replication rate (after aggregating across all 6344 participants) to argue that reproducibility is high when extremely high power is used. This interpretation is based on ML2014's small, ad hoc sample of classic and new findings, as opposed to OSC-2015's effort to examine a more representative sample of studies in high-impact journals. Had GKPW selected the similar Many Labs 3 study (5) they would have arrived at a more pessimistic conclusion: a 30% overall replication success rate with a multi-site, very high-powered design.*

The authors of OSC-Reply miss the point of our analysis. We asked the question, "When the replication rate is known to be high, as it was in ML2014, what would happen if one instead used the low-powered design used by OSC-2015?" The answer, as we demonstrated, is that the replication rate would drop precipitously. In other words, we used a data set with a known high replication rate to estimate the "cost" of using the low powered methods used in OSC-2015. The cost was substantial.

The authors of OSC-Reply go on to say:

> *That said, GKPW's analysis demonstrates that differences between labs and sample populations reduce reproducibility according to the CI measure. Also, some true effects may exist even among non-significant replications (our additional analysis finding evidence for these effects is available at https://osf.io/smjge/). True effects can fail to be detected because power calculations for replication studies are based on effect sizes in original studies. As OSC-2015 demonstrates, original study effect sizes are likely inflated due to publication bias.*

This argument hinges on OSC-2015's conclusion that a large proportion of psychological findings are not replicable, and thus the effect sizes that are reported in the original studies are inflated. As OSC-Reply concedes, there is not enough evidence in their data to draw "pessimistic conclusions about reproducibility." Therefore, they provide no evidence that the effect sizes in original studies were inflated, nor do they provide an estimate of publication bias.

Their next argument is:

> *Unfortunately, GKPW's focus on the CI measure of reproducibility neither addresses nor can account for the facts that the OSC-2015 replication effect sizes were about half the size of the original studies on average, and 83% of replications elicited smaller effect sizes than the original studies. The combined results of OSC-2015's five indicators of reproducibility suggest that even if true, most effects are likely to be smaller than the original results suggest.*

As we already noted, the replication studies may have had smaller effect sizes because in many cases their procedures and populations departed significantly from those of the original studies. Also, we will note once again that neither we nor the authors of OSC-2015 found any substantive differences in the conclusions drawn from the confidence interval measure versus the other measures.

In summary, none of the arguments made in OSC-Reply disputes the fact that the authors of OSC-2015 used a low-powered design, and that (as our analyses of the ML2014 data demonstrate) this likely led to a gross underestimation of the true replication rate in their data.

Bias

Many of the OSC-2015 replication studies differed markedly from the original studies in ways that made successful replication less likely. To estimate the fidelity of the replication studies, we used OSC-2015's measure of whether the authors of the original study had endorsed the methodological protocol of the replication study, and we found that endorsed studies were four times more likely to replicate than were unendorsed studies.

The authors of OSC-Reply raise a number of objections to this finding. For example, they point out that one of the six low fidelity studies that we mentioned in our Technical Comment was in fact replicated successfully. First, one-in-six is quite consistent with our claim that low fidelity

studies were unlikely to replicate. Second and more importantly, the six studies we mentioned were merely meant to provide a few easy-to-understand examples and were not an exhaustive list. There are many more cases of replication studies that differed significantly from the original studies (and we will describe another particularly striking example below).

The authors of OSC-Reply go on to criticize the measure of fidelity we employed:

> *There is an alternative explanation for the correlation between endorsement and replication success; authors who were less confident of their study's robustness may have been less likely to endorse the replications. . . In sum, GKPW made a causal interpretation for OSC-2015's reproducibility with selective interpretation of correlational data.*

We were very clear in our Technical Comment that endorsement by the original investigators is an imperfect indicator of fidelity, and in fact, we explicitly raised the very alternative interpretation that the authors of OSC-Reply raise in the comment above. It may well be that some of the original authors failed to endorse the replication protocols because the original authors lacked confidence in their own results. But it is worth noting that this is not the interpretation that the authors of OSC-2015 offered to their readers. They reported that some authors failed to endorse because they "maintained concerns based on informed judgment/speculation" or "on published empirical evidence for constraints on the effect" (Supplementary Materials, p. 44).

The authors of OSC-2015Reply then offer a different measure of fidelity:

> *In fact, OSC-2015 tested whether rated similarity of the replication and original study was correlated with replication success and observed weak relationships across reproducibility indicators (e.g., r = .015 with "p < .05" criterion, SI, p. 67, https://osf.io/k9rnd).*

However, the authors of OSC-Reply omit one important detail about this measure, namely, that the ratings of replication similarity to which they refer were *made by the researchers who conducted the replications* and who may have had their own reasons for rating their studies as high or low in fidelity. Surely the original researchers who conceived and conducted the original study, often as part of a long program of research, are better able to determine what constitutes a high fidelity replication than are replicators who almost always had less (if any) expertise and familiarity with the research paradigms.

In short, the ratings of fidelity by the original investigators and the replicators are both imperfect, but the former are more likely to be reliable because they are made by experts. A few examples will serve to illustrate this point. In one original study, researchers asked Israelis to imagine the consequences of taking a leave from mandatory military service (Shnabel & Nadler, 2008). The replication study asked Americans to imagine the consequences of a taking a leave to get married and go on a honeymoon. Not surprisingly, the original authors expressed "concerns based on informed judgment/speculation" and did not endorse the replication study. And not surprisingly, the replication study failed—and yet, the replicator rated the replication as "extremely similar" to the original protocol, which was the second-highest rating of fidelity.

In another study, White students at Stanford University watched a video of four other Stanford students discussing admissions policies at their university (Crosby, Monin, & Richardson, 2008). Three of the discussants were White and one was Black. During the discussion, one of the White students made offensive comments about affirmative action, and the researchers found that the observers looked significantly longer at the Black student when they believed he could hear the others' comments than when he could not. Although the participants in the replication study were students at the University of Amsterdam, they watched the same video of Stanford students talking (in English!) about Stanford's admissions policies. In other words, unlike the participants in the original study, participants in the replication study watched students at a foreign university speaking in a foreign language about an issue of no relevance to them. The replicators acknowledged that these were important differences that might well have influenced the results: "The cultural background of the participants is totally different. The original effect builds on a typical US university admission selection, highly competitive. Here we do not have such a procedure." What did the original authors think about the replication protocol? They naturally expressed concerns and did not endorse it. And not surprisingly, the original study done at Stanford did not replicate at the University of Amsterdam—and yet, for some reason, the replicators (who had explicitly acknowledged the differences between the original study and the replication study) gave the two studies the highest rating of "virtually identical."

To the replicators' credit, they recognized that all of this might matter, so they ran two other versions of the study at a small college in Massachusetts, one in which participants saw the Stanford video and another in which they saw this video with all references to Stanford edited out. The latter version nearly replicated the original findings (predicted interaction: $p = .078$, key simple effect, $p = .033$). In other words, the replicators conducted a study that differed in key ways from the original and found that the original study did not replicate, and then showed that when these differences were eliminated the original study did replicate. And yet, OSC-2015 included *only the failure to replicate conducted at the University of Amsterdam* in their results. Anyone who carefully reads all the replication reports will find more troubling examples like these.

Lastly, OSC-Reply made these arguments:

> *Consistent with the alternative account, prediction markets administered on OSC-2015 studies showed that it is possible to predict replication failure in advance based on a brief description of the original finding (6). Finally, GKPW ignored correlational evidence in OSC-2015 countering their interpretation such as evidence that surprising or more underpowered research designs (e.g., interaction tests) were less likely to replicate.*

These results have no bearing on the question of the fidelity of the *methods* of the replication studies.

SAMPLING ISSUES

And now for the bad news. Even if the authors of OSC-2015 had used a *perfect* method for taking *every* source of error into account, and even if they had conducted nothing but high fidelity replications, their data would still be fundamentally incapable of answering the main question that they set out to answer. Why?

The authors of OSC-2015 stated that their goal was "to obtain an initial estimate of the reproducibility of psychological science." Indeed, these words are even in the title of their article. Now, as every scientist, pollster, and college student knows, when one wishes to use a sample to estimate a parameter of a population, then the sample must either be (a) randomly selected from the population or (b) statistically corrected for the bias that non-random selection introduces. The authors of OSC-2015 did neither of these things. Instead, they used non-random selection procedures that made it extremely likely that the studies they chose to replicate were unrepresentative of the literature about which they wished to make an inference. Here is what they did.

The population of interest was the research literature of psychological science. The authors of OSC-2015 began by defining a non-random sample of that population: the 488 articles published in 2008 in just three journals that represented just two of psychology's subdisciplines (i.e., social psychology and cognitive psychology) and no others (e.g., neuroscience, comparative psychology, developmental psychology, clinical psychology, industrial-organizational psychology, educational psychology, etc.). They then reduced this non-random sample not by randomly selecting articles from it, but by applying an elaborate set of selection rules to determine whether each of the studies was eligible for replication—rules such as "If a published article described more than one study, then only the last study is eligible" and "If a study was difficult or time-consuming to perform, it is ineligible" and so on. Not only did these rules make a full 77% the non-random sample ineligible, but they also further biased the sample in ways that were highly likely to influence reproducibility (e.g. researchers may present their strongest data first; researchers who use time-consuming methods may produce stronger data than those who do not; etc.). After making a list of selection rules, the authors of OSC-2015 permitted their replication teams to break them (e.g., 16% of the replications broke the "last study" rule, 2 studies were replicated twice, etc.).

Then it got worse. Instead of randomly assigning the remaining articles in their non-random sample to different replication teams, the authors of OSC-2015 invited particular teams to replicate particular studies or they allowed the teams to select the studies they wished to replicate. Not only did this reduce the non-random sample further (a full 30% of the articles in the already-reduced sample were never accepted or selected by a team), but it opened the door to exactly the kinds of biases psychologists study, such as the tendency for teams to accept or select (either consciously or unconsciously) the very studies that they thought were least likely to replicate. As the authors of OSC-Reply remind us, even casual bystanders in a prediction market can tell beforehand which effects will and will not replicate—and yet, armed with exactly those insights, replication teams were given a *choice* about which studies they would replicate. Then, in a final blow to the notion of random sampling, the already reduced non-random sample was non-randomly reduced one last time by the fact that 18% of the replications were never completed.

If this same procedure had been used to estimate a parameter of a human population rather than of a scientific field, no reputable scientific journal would have published the findings. And indeed, the authors of OSC-2015 acknowledged that the cumulative effect of their repeated non-random selection and re-selection was a problem *so serious* that it made their goal of parameter-estimation impossible. Referring to their own findings, they concluded, "It is unknown the extent to which these findings extend to the rest of psychology" because "the impact of selection bias is unknown." This is an admirably candid but truly remarkable conclusion that many readers of OSC-2015 seem to have overlooked. As the authors of OSC-2015 themselves admit, the failure to follow standard practice with regard to sampling a population means that *their findings cannot be used to assess the reproducibility of psychological science*. Given that this was the primary goal of their project, it is difficult to see what value their findings have.

SUMMARY

(1) The authors of OSC-2015 conducted (and then interpreted the results of) many low fidelity replications, even though there is clear evidence that these low fidelity replications were much more likely to fail.

(2) The authors of OSC-2015 did not measure the error introduced by infidelity, and therefore incorrectly estimated how many of their studies should have failed by chance alone.

(3) The authors of OSC-2015 used a low-powered design (i.e., one replication per study), and there is clear evidence that this led them to drastically underestimate the true replication rate in their own data.

(4) The authors of OSC-2015 non-randomly selected and re-selected the original studies that they attempted to replicate, and therefore never had any chance of obtaining "an initial estimate of the reproducibility of psychological science."

(5) OSC-2015 provides no evidence of a replication crisis in psychological science.

REFERENCES

Crosby, J. R., Monin, B., & Richardson, D. (2008). Where do we look during potentially offensive behavior? *Psychological Science*, *19*, 226-228.

Shnabel, N. & Nadler, A. (2008). A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting reconciliation. *Journal of Personality and Social Psychology*, *94*, 116-132.