On the Importance of Modeling the Invisible World of Underlying Effect Sizes

Brent M. Wilson and John T. Wixted

Department of Psychology, University of California, San Diego,

Author note

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Brent M. Wilson or John T. Wixted, Department of Psychology, University of California, San Diego, La Jolla, California 92093, USA

E-mail: b6wilson@ucsd.edu

E-mail: jwixted@ucsd.edu

Abstract

The headline findings from the Open Science Collaboration (2015)—namely, that 36% of original experiments replicated at $p < .05$, with the overall replication effect sizes being half as large as the original effects—cannot be meaningfully interpreted absent a formal model. A simple model-based approach might ask: what would the state of original science be and what would replication results show if original experiments tested true effects half the time (prior odds = 1), true effects had a medium effect size (Cohen's δ = 0.50), and power to detect true effects was 50%? Assuming no questionable research practices, 91% of $p < .05$ findings in the original literature would be true positives. However, only 58% of original $p < .05$ findings would be expected to replicate using the Open Science Collaboration approach, and the replication effects overall would be only ~60% as large as the original effects. A minor variant of this model yields an expected replication rate of only 45%, with overall replication effect sizes dropping by half. If the state of original science is as grim as a non-model-based (i.e., intuitive) interpretation of the Open Science Collaboration data suggests, should it be this easy to largely account for those findings using a model in which 91% of statistically significant findings in the original science literature are true positives? Claims that the findings reported by the Open Science Collaboration indicate a replication crisis should not be based solely on intuition but should instead be accompanied by a specific model that supports that interpretation.

Keywords: Null hypothesis significance testing; False Positives; Positive predictive value; Replication crisis

Introduction

The reliability of psychological science is often evaluated in terms of true positives and false positives. A true positive is defined as a $p < .05$ finding associated with an underlying effect size that differs from 0 in the same direction as the measured effect size, whereas a false positive is defined as a $p < .05$ finding with an underlying effect size equal to 0 despite the significant outcome. Ideally, positive predictive value (PPV), which is the proportion of $p < .05$ findings that are true positives, would be high. However, in recent years, a surprisingly large proportion of experiments, when conducted again by independent labs, failed to replicate at $p < .05$. Such findings have been interpreted to mean that many ostensible discoveries in the experimental psychology literature were in fact false positives, so PPV is actually low.

This unexpected revelation is often conceptualized in terms of a replication crisis, one that, in theory, can be ameliorated by reforming scientific practices. Some of the most popular reforms include preregistering experiments (Nosek, Ebersole, DeHaven, & Mellor, 2018), conducting experiments with higher power (Button et al., 2013), eliminating publication bias (Bishop, 2019; van Assen, van Aert, Nuijten, & Wicherts, 2014), and making the relevant data and materials publicly available on repositories such as OSF.

We do not fully embrace this perspective because, in our view, the problem has been partially misdiagnosed. According to our perspective, the misdiagnosis is based on a possible misinterpretation of a large-scale effort to replicate 100 representative experiments from the fields of cognitive and social psychology (Open Science Collaboration, 2015, henceforth OSC2015). This is the only large-scale study that attempted to assess the reliability of psychological science broadly considered. As they put it: “In this Research Article, we report a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of

psychological science" (p. aac4716-1). Their conscious effort to minimize selection bias differs from other replication studies that focused on selected—often influential and/or seemingly implausible—findings. Examples of the latter include replications of findings purporting to show that ESP is real (Ritchie, Wiseman, & French, 2012) or that simply drawing a line between close or distant points led participants to feel correspondingly close or distant to their family members (Pashler, Coburn, & Harris, 2012).

In the metascience literature, discussions of efforts to replicate representative findings vs. selected findings are often conflated, as if they tell one story. In our view, they tell two different stories, both of which need to be understood to effectively address the actual problem. In our view, the actual problem was revealed by a series of unexpected failures to replicate influential (selected) findings when independent labs conducted direct replications with high power (e.g., Harris et al., 2013; Klein et al., 2018; Pashler et al., 2012; Ritchie et al., 2012; Rohrer et al., 2015). How did the field come to accept those influential findings as true positives, and what can be done to prevent that from happening going forward?

This is where we believe the focus should be—that is, on findings that have been selected precisely because they are influential and especially if they seem implausible (Wilson et al., 2020, 2022). Placing the focus on influential findings would address the primary problem in a much more cost-effective way compared to trying to ensure that every published experiment, no matter how obscure, is conducted with high power and is also replicated by an independent lab to ensure its validity (Lewandowsky & Oberauer, 2020). This would be especially true if the original science literature is already in better shape than meets the eye.

We realize, of course, that many experimental psychologists (perhaps most) have a different view according to which psychological science is full of false positives and is therefore

in need of serious methodological improvement. This perspective is largely informed by the results of OSC2015. Of the 100 experiments they replicated, 97 originally reported statistically significant ($p < .05$) findings. However, when tested again, only 36% of them replicated at $p < .05$, as if only 36% of statistically significant findings in psychological science are true positives (i.e., as if PPV = .36). Ioannidis (2015) succinctly summarized this common interpretation by declaring that "…nearly two-thirds of the original findings were false-positives" (p. 4.). Similar sentiments can be found in many other articles (e.g., Flake et al., 2022; Malich & Rehmann-Sutter, 2022; Owens, 2018; Sikorski, 2022; Scheel, 2022). If correct, this interpretation would mean that the replication problem is not limited to selected experiments that report influential (often implausible) results but is instead pervasive.

However, we contend that the simplest and most straightforward model-based interpretation of the OSC2015 findings suggests the opposite (Wilson et al., 2020). That is, by a considerable margin, most findings in the original science literature are true positives, not false positives. Why, then, does it seem otherwise to so many? We think it is because, almost without exception, the interpretation of OSC2015 has been based on intuition, as if the implications of the 36% replication rate are self-evident. Yet it seems safe to suggest that intuition can be misleading, particularly in the domain of statistics. Therefore, a formal model—even a very simple one—should always be used to interpret the replication data. So far as we know, no one who interprets the OSC2015 as being indicative of a replication crisis has done that, and the main purpose of our article is to suggest that, going forward, they do.

To illustrate the importance of modeling replication data, we consider two simple models of underlying reality, with one being a slight variant of the other. Both models interpret the OSC2015 data to mean that PPV in the original literature is high. Before presenting those

models, it seems important to first address a common response we have encountered when suggesting that the results of OSC2015 might not be what they seem to be at first glance. The response is that we are therefore suggesting that everything is fine with psychological science, as if we are denying a reality that is plainly apparent to those who are willing to see it. However, we are not suggesting that everything is fine with psychological science. We are instead suggesting that there is something *seriously wrong* with psychological science, but the problem may have been misdiagnosed. If so, the proposed solutions may be correspondingly off the mark. From our perspective, the problem has nothing to do with the results of OSC2015, which replicated representative findings in psychological science, and everything to do with the results of replication studies that attempted to reproduce selected (usually influential and often implausible) findings that were previously thought to be true positives.

Modeling the OSC2015 Replication Results

Intuitively, if an original $p < .05$ finding is a true positive, it seems reasonable to expect that a replication experiment would again be statistically significant with an effect size close to the one that was originally reported. Indeed, it *is* intuitive, and therein lies the problem. According to Table 1 of OSC2015, the replication experiments had 92% power to detect an effect size *as large as the one reported in the original experiment*, not 92% power to detect the true underlying effect size. However, and critically, unless the original experiments were conducted with ~100% power (which is inconceivable), the observed effect sizes for the 97 experiments reporting statistically significant results were inflated relative to their true underlying effect sizes. If so, the replication experiments necessarily had lower power than advertised, and the replication effect sizes would, on average, have to be smaller than the originally reported effect sizes (Maxwell et al., 2015).

Why are observed statistically significant effect sizes (e.g., observed Cohen's d values) inflated relative to the true underlying effect size (e.g., Cohen's δ) in the absence of 100% power? The reason is that observed effect sizes will vary from experiment to experiment due to measurement error. With 50% power to detect an underlying effect size of $\delta = 0.50$, for example, the 50% of observed effects greater than 0.50 will be statistically significant, whereas the 50% of observed effects smaller than 0.50 will not be.[1] If we focus only on the statistically significant Cohen's d values, all of which exceed 0.50, the average of them will also obviously exceed 0.50. In other words, the average of the statistically significant effect sizes will be inflated relative to the underlying effect size. This is a statistical reality that is not related to publication bias, per se. Publication bias obviously exists, but statistically significant effect sizes will be inflated whether they are published in a journal or posted to a preprint server.

How inflated should we expect the original effect sizes replicated by OSC2015 to be, and how much should we expect them to decrease upon replication? In addition, what percentage of the original experiments reporting $p < .05$ results should we expect to replicate at $p < .05$? Intuition cannot answer these questions. Instead, a model is needed, one that takes into account the effect-size inflation associated with $p < .05$ results.

The minimal assumptions needed to model the results reported by OSC2015 consist of (1) an assumption about power in the original experiments, (2) an assumption about the true underlying effect sizes of the original experiments, and (3) an assumption about the prior probability of testing a true (non-zero) effect in the original experiments. For our first model, we tried to adopt principled and defensible assumptions about these three issues.

[1] The 50th percentile of observed effect sizes would actually differ somewhat from .50 because the distribution of Cohen's d is somewhat skewed.

Model I

Previous work suggests that experiments in the field of psychology have about ~50% power to detect a medium effect size (e.g., Cohen, 1962; Fraley & Vazire, 2014; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989; Szucs & Ioannidis, 2017). We therefore adopt these two assumptions (i.e., 50% power to detect a Cohen's δ of 0.50). Note that the average replication effect size of OSC2015 was $\bar{d} \approx 0.50$, and this is a relatively unbiased estimate of the true average underlying effect size associated with the original experiments in psychology. Thus, these assumptions seem to us to be more defensible than any other assumptions one might make about power and effect sizes in the original psychological science literature.

With only one more assumption about the base rate of testing true effects in original science, we would have a model of underlying reality that makes predictions and can therefore be evaluated in relation to the replication results reported by OSC2015. The true base rate is not knowable, so we make an assumption about its value that is maximally noncommittal to unknown information. Specifically, we assume that when experimental psychologists conduct experiments, true and false effects are tested with equal probability (i.e., the base rate is .50, which is to say that the prior odds are 1/1 = 1).[2]

Model I assumes no questionable research practices (QRPs). As we define them, QRPs are actions that a scientist inappropriately takes when conducting an experiment or analyzing data (e.g., p-hacking, secretly excluding inconvenient outliers, etc.). According to that definition, publication bias is not a QRP. Instead, it is an editorial policy governing the publication of research articles. There are obvious costs associated with publication bias, but we have argued

[2] The null hypothesis of no difference is never strictly true (Wilson et al., 2022), but here we adopt standard null-hypothesis significant testing for simplicity. Also, Rouder et al. (2009) noted that "…it is often natural to set the prior odds to 1.0, a position that favors neither the null nor the alternative" (p. 228).

that there are clear benefits as well (Wilson et al., 2020; 2022). For example, Wilson et al. (2020) argued that publishing the file drawer of non-significant replications that everyone one would like to see would mean also publishing the (possibly much larger) junk drawer of useless and uninformative results, degrading the scientific literature overall. In addition, Wilson et al. (2022) argued that PPV at the level of theory can be maximized by using an optimal sample size (not too small, but not too large either) while publishing claims associated with $p < .05$ results that confirm a theory-based prediction.

Different researchers will have different cost-benefit analyses associated with publication bias, and some (like us) will conclude that the benefits of publication bias outweigh the costs. We realize that many disagree, but our only point is that Model I assumes no QRPs as we define them. For simplicity, the model also assumes complete publication bias (i.e., only $p < .05$ results are published). This seems like a reasonable assumption given that OSC2015 attempted to select representative findings from psychological science and found that 97% reported statistically significant (positive) results. Similarly, Scheel, Schiken, and Lakens (2021) recently found a 96% positive result rate in standard reports in psychological science.

What does this simple model of underlying reality (power $= .50, \delta|H0 = 0, \delta|H1 = 0.50, P(H_0) = P(H_1) = .50$) predict about PPV in the original science literature? And what does it predict about the degree to which statistically significant effect sizes would be inflated? PPV in the original science literature would be $\frac{(1-\beta)P(H_1)}{[(1-\beta)P(H_1)+\alpha P(H_0)]} = \frac{.50\times.50}{.50\times.50 + .05\times.50} = \frac{.25}{.25 + .025} = .91$. In other words, 91% of statistically significant results in the original science literature would be true positives and only 9% would be false positives. In addition, a large number of simulation trials based on Model I reveals that the expected statistically significant effect size in the original science literature would be approximately $\bar{d} = 0.79$. This value is considerably

inflated above the true underlying effect size of 0.50, but it is smaller than the average effect size associated with the original findings replicated by OSC2015 ($\bar{d} = 0.99$).

If we then set out to replicate a representative subset of those significant results, 91% would be true positives with an underlying effect size equal to 0.50, and 9% would be false positives with an underlying effect size of 0. With 92% power to detect the inflated effect size of 0.79, the replication experiments would actually have only 63% power to detect the $\delta = 0.50$ effect size associated with the true positives. As shown in the top row of Table 1, only 58% of the original experiments reporting $p < .05$ results would be expected to replicate at $p < .05$ (lower than expected based on 63% power because some of the replications test false positives). The expected replication effect size for the same experimental protocol conducted with a new random sample of subjects would not be the inflated effect size of 0.79 but would instead be ~0.48 (~60% of the original inflated effect size). Once again, the expected effect size is less than 0.50 because 9% of the replication experiments test false positives with effect sizes of 0. The predicted replication effect size of 0.48 is close to the average replication effect size reported by OSC2015, namely $\bar{d}$ = 0.50.

In the OSC2015 data, only 36% of original experiments replicated at $p < .05$, whereas this simple model predicts a 58% replication rate given methodologically flawless original experiments (i.e., no QRPs) and methodologically flawless replications (i.e., no methodological infidelities). Thus, there is more to the story than what this simple model predicts. The rest of the story might consist of QRPs in the original experiments or methodologically less-than-perfect replication experiments (or, perhaps most likely, some combination of the two). Still, it seems important to appreciate how much of what OSC2015 found is anticipated by this simple model. Comparing the 36% replication rate and 50% reduction in effect size to an intuitively expected

92% replication rate and undiminished effect size seems compatible with a crisis interpretation. Comparing the OSC2015 results to a model-based expected 58% replication rate and 40% reduction in effect size seems much less compatible with a crisis interpretation.

Table 1. Predicted and observed data assuming 50% power for the original experiments.

	Original Experiments				Replication Experiments			
source	$p(H_1)$	$\delta \mid H_1$	$\bar{d} \mid p < .05$	PPV	$p(H_1)$	$\delta \mid H_1$	$\bar{d}$	p(rep)
Model I	0.50	0.50	0.79	0.91	0.91	0.50	0.48	0.58
Model II	0.50	0.60	0.98	0.91	0.91	0.50	0.49	0.45
OSC2015	??	??	0.99	??	??	??	0.50	0.36

Note. For the original experiments, $p(H_1)$ is the probability that the alternative hypothesis is true in the set of experiments conducted by experimental psychologists, $\delta \mid H_1$ is Cohen's δ (the true underlying effect size) given H_1, $\bar{d} \mid p < .05$ is the average Cohen's d for statistically significant results, and PPV represents positive predictive value (i.e., the proportion of $p < .05$ findings that are true positives). For the replication experiments, p(H_1) is the probability that the alternative hypothesis is true in the set of replications conducted, $\bar{d}$ is the average of the replication effect sizes, and p(rep) is the probability of obtaining a statistically significant replication. For OSC2015, ?? means that the information is unknown.

Model II

So far, our analysis in terms of a simple model has assumed that the replication experiments were *exact* replicas of the original experiments, with the only difference being that they were conducted using a different random sample of subjects. Obviously, exact replications are not possible in many cases (e.g., the original authors might leave out details about the materials used), and the inevitable noise introduced by translating the original experimental protocol into the replication experimental protocol might reduce the underlying effect size associated with the replication experiments below the original effect size. This would be true even if the replication experiments are methodologically flawless given the information available from the article reporting the original findings.

It might seem as though the inevitable methodological deviations from the original experiment would be as likely to increase as decrease the underlying effect size. That might be

true had the original effects not been selected using a $p < .05$ filter, but because they were, we think the methodological deviations from the original experiments would be more likely to decrease the effect size. For example, imagine that a random sample of stimuli used in an original experimental protocol, such as a random sample of words from the dictionary, positively influenced the outcome of the experiment for theoretically uninteresting reasons (cf. Baribault et al., 2018). Because the particular words are theoretically irrelevant, they would not matter to the original scientist and might not be reported in the paper. If a perfectly valid replication experiment used a different random selection of stimuli from the same source, the true underlying effect size of the replication protocol would be lower than that of the original protocol.

Again, this is because the original effect was selected using a $p < .05$ filter, and any random benefit from the original stimuli would be lost upon replication. As an example, according to the OSC2015 supplementary materials, one replication experiment was characterized as differing from the original experiment in the following way: "Generating the materials (symbol sequences) involves randomness, therefore the original and the replication materials are not identical (original materials could not be obtained)."

In making this point, we are not arguing that the replication experiments in OSC2015 were methodologically flawed. Whether or not they might have been flawed is a separate debate (e.g., Gilbert et al., 2016), one that we sidestep by assuming that the replication experiments were methodologically flawless. Our point is that even a methodologically flawless replication experiment might have a slightly smaller underlying effect size than the corresponding original experiment because the original finding was selected using a $p < .05$ filter.

We know of no principled way to estimate the magnitude of this effect, but because any such effect would presumably be small, we set it equal to half of what is usually considered to be a small effect size in terms of Cohen's d (i.e., half of 0.20, or 0.10). We implement this assumption in Model II by adding 0.10 to the assumed effect size associated with the original experiments, bringing it to 0.60. We could instead subtract it from the expected replication effect size of 0.50, but the results of OSC2015 provide an unbiased estimate of the true average underlying effect size associated with the replication experiments, and it came to 0.50. It therefore makes sense to keep that estimate for the replication experiments and to include this methodological consideration by adding 0.10 to the (unknown) original effect size.

If the original experiments had 50% power to detect a true underlying effect size of 0.60, the inflated effect size associated with statistically significant outcomes would be $d = 0.98$, which is now almost identical to the average of observed original effect sizes in OSC2015 ($d = 0.99$). If we then set out to replicate a representative subset of those significant results, 91% would still be true positives with an underlying effect size equal to 0.50, and 9% would be false positives with an underlying effect size of 0. With 92% power to detect the inflated effect size of 0.98, we would actually have only 49% power to detect the 0.50 effect size associated with the true positives. Altogether, only 45% of the original experiments reporting $p < .05$ results would be expected to replicate at $p < .05$. The expected replication effect size would be ~0.49, which is nearly identical to the average replication effect size reported by OSC2015 of 0.50. These results are summarized in the middle row of Table 1. The observed 36% replication rate in OSC2015 is still somewhat lower than the expected value of 45%, but not by much.

The OSC2015 Replication Failures

Critically, both Model I and Model II assume that many of the "replication failures" in OSC2015 are mostly successful replications that did not achieve $p < .05$ because the replication experiments were underpowered. The predicted average replication effect sizes for the non-significant replications, though small (0.20 in Model I, and 0.23 in Model II), are greater than zero (left column of Table 2). Both models also predict that a substantial proportion of the nonsignificant replications will have effect sizes going in the same direction as the originally reported effect size, namely, 85% for both (right column of Table 2). These models predict that such trends should be observed in the OSC2015 replication data if we look for them. By contrast, if the non-significant replications in OSC2015 are false positives, as is commonly assumed, the expected average effect size for the non-significant results would be $\bar{d} \approx 0$, with only ~50% being in the same direction as the original experiments.

Wilson et al. (2020) analyzed the non-significant replications from OSC2015 and found that the average effect size was $\bar{d} = 0.14$, an outcome significantly greater than 0 at $p < .001$. Similarly, a Bayesian analysis performed on these data strongly favors the alternative hypothesis over the point-null hypothesis, $B_{10} = 27.5$. In addition, 73% of the non-significant replication effect sizes went in the same direction as the original effect size (only 27% in the opposite direction), which is significantly greater than the 50% value that would be expected if they were all statistical false positives ($p = .002$). These results are summarized in the bottom row of Table 2. The models overpredict the magnitude of these effects, perhaps reflecting QRPs in the original experiments or methodological infidelities in the replication experiments. Either way, it seems noteworthy that a simple model that predicts a PPV of .91 in the original science literature can accommodate much of the OSC2015 replication results.

Table 2. Trends associated with non-significant replications in OSC2015.

source	$\bar{d}\|p > .05$	$p\,(d > 0)$
Model I	0.20	0.85
Model II	0.23	0.85
OSC2015	0.14	0.73

Note. Here, $d \mid p > .05$ is the average d for non-significant results, and p($d > 0$) is the probability that a non-significant replication will have an effect size in the same direction as the original experiment.

Conclusion

Our modeling effort is intended to illustrate why the interpretation of OSC2015 as an impeachment of psychological science—as opposed to an endorsement of it—requires the specification of a formal model. This contrasts with the almost universal belief that intuition alone is up to the task. For example, in a recent editorial, Renkewitz and Heene (2019) summarized the OSC2015 findings in a way that has become quite common: "With estimated replication rates ranging between 25% for social psychology and 50% for cognitive psychology (Open Science Collaboration, 2015), it became obvious that psychology suffers from a severe replicability problem." Our point is that it is *not* obvious, intuition notwithstanding.

The models we considered here were designed to be noncommittal with respect to unknown information, not to produce a preferred interpretation. Are effect sizes in experimental psychology typically small, medium, or large? The answer is unknown, so we assumed a medium effect size of $d = 0.50$. Is power to detect a medium effect size 20%, 50%, or 80%? Again, the answer is unknown, but it has long been thought to be approximately 50%, so we chose that value. Is the base rate of testing a true effect .20, .50, or .80? Yet again, the answer is unknown, so we chose .50. That is literally all there is to Model I, and the expected replication rate assuming no QRPs in the original experiments and no methodological infidelities in the replication experiments is only 58%, with the overall replication effect sizes being only 61% of the originally reported effect sizes.

Allowing for the possibility that the replication experiments have slightly smaller underlying effect sizes than the original experiments due to stimulus selection effects brings the expected replication rate to only 45%, with the overall replication effect sizes being only 50% of the originally reported effect sizes. Relative to these models, QRPs in the original experiments and methodological infidelities in the replication experiments do not have a lot to add to our understanding of OSC2015. Although this conclusion contrasts with a widespread perspective, it accords with a much more complex and more realistic model of underlying reality (involving a distribution of underlying effect sizes and a distribution of power) described by Wilson et al. (2020).

What formal model of underlying reality suggests otherwise? The main point of our paper is not that our models are necessarily the best guides to use. Instead, the main point is that those who interpret the results of OSC2015 in terms of a replication crisis have relied on intuition instead of specifying the model they have in mind and then showing how it supports that interpretation in light of the relevant empirical evidence. How do you know that the results imply a replication crisis without basing that interpretation on a formal model that makes predictions about what should be observed? Evaluating the 36% replication rate reported by OSC2015 against an intuitive ideal that is not specified should be discontinued in favor of specifying a simple formal model that supports that intuitive interpretation and explaining why it should be preferred to the simple models we considered here.

It might seem as though prior modeling efforts have already addressed this issue. For example, in an influential paper entitled “Why most published research findings are false,” Ioannidis (2005) used equations combined with different assumptions about base rates, power, and various biases to show that most scientific findings could be false. He then expressed the

opinion that most findings are in fact false. However, nowhere in that paper did he attempt to reconcile his negative evaluation with a model-based interpretation of actual replication data. Instead, the paper shows what could be true given the right assumptions. Other assumptions would lead to a different conclusion, so it is important to ask which assumptions better accommodate empirical replication data.

In another influential paper, Simmons, Nelson, and Simonsohn (2011) showed how easy it would be for psychologists to publish statistically significant false positives by engaging in QRPs like *p*-hacking. It certainly would be easy to do that, but there was no effort in that paper to relate the quantitative simulations to actual replication data to determine how much it might contribute to an understanding of the results. There is little doubt that scientists sometimes engage in QRPs. After all, they admit it when asked (John, Loewenstein, & Prelec, 2012). However, our main question of interest is not whether QRPs sometimes occur but is instead how much of the replication results they explain. That question cannot be answered by relying on intuition, and the formal models considered here suggest that whatever role they play might be relatively small (cf., Head et al., 2015).

The influential papers discussed above were published at a time of growing concern about the reliability of scientific research but before the results of OSC2015 were published. When OSC2015 was published, it appeared to provide confirmation of the concerns that had been raised in these earlier reports. However, absent a formal model, it is not clear if the OSC2015 results should confirm those fears or allay them.

Our model-based analysis is in some ways consistent with a Bayesian analysis of the OSC2015 findings conducted by Etz and Vandekerckhove (2016). They found that none of the replication OSC2015 experiments provided strong evidence in favor of the null, an interpretation

that is diametrically opposed to the intuition-based view that two-thirds of findings in experimental psychology are false positives (e.g., Ioannidis, 2015). Then again, Etz and Vandekerckhove (2016) also found that most of the experiments considered by OSC2015 (64%) did not provide compelling evidence for either the null or the alternative hypothesis, in either the original or replication experiments. Thus, their main message is that both the original findings and the OSC2015 findings are largely inconclusive. Even if that is true, citing the results of OSC2015 as providing compelling evidence of a replication crisis would still be premature.

If either Model I or Model II provide an accurate depiction of underlying reality, then efforts to reform psychological science in light of OSC2015 may not be as urgent as many appear to believe. Critically, this is not to say that all is well in experimental psychology. Instead, a serious problem exists given that influential findings that were once thought to be true positives turned out to be false positives when replicated with high power by an independent lab. In our view, *that* is the main problem, which is why we believe the focus should be placed on independent labs conducting large-N direct replications of influential experiments (Wilson, 2020, 2022). Our perspective seems compatible with Isager et al.'s (2021) suggestion that, given limited resources, experiments should be accepted for replication if they are deemed to have high replication value in terms of expected utility gain.

Camerer et al. (2018) found that people are surprisingly good at predicting which experiments will replicate and which will not. This suggests that influential experiments that fail to replicate have detectably low prior odds of being true (e.g., ESP is real), not the even odds assumed by Models I and II for more representative experiments. An experiment with low prior odds of being true that nevertheless achieves $p < .05$ should increase one's belief that the result is a true positive but not necessarily to the point of believing that it is more likely than not to be

true. For example, in the case of an experiment purporting to show that ESP is real ($p < .05$), one's belief that it reflects a true positive might increase from extremely unlikely before the experiment was conducted to very unlikely after it was conducted. But because it remains very unlikely to be true, the expectation should be that it should fail to replicate despite achieving $p < .05$ in the original experiment (Wilson & Wixted, 2018). The surprise would be if it successfully replicated.

Implausible findings that originally achieve a $p < .05$ outcome often become influential because, despite being implausible, they appear to have the scientific stamp of approval. As a result, implausible (and likely non-replicable) findings tend to be overrepresented in the relatively small set of findings that become influential. This is why we believe that replication science should focus on validating (or not) influential findings (Wilson & Wixted, 2020, 2022). It is not only a cost-effective approach, but it also targets the main problem because, at least according to the models we considered here, psychological science may be in better shape than many believe in light of OSC2015.

And here, we find some common ground with those who interpret OSC2015 in a different way than we do. For example, despite interpreting the OSC2015 data as indicating that nearly two-thirds of statistically significant findings in psychological science are false positives, Ioannidis (2015) also had this to say: "For the more influential and heavily cited studies, the imperative for independent exact replication should be very hard to resist, these studies should be subjected to replication. It would make little sense to neglect to replicate a study upon which hundreds or thousands of other investigations depend" (p. 8).

At least on this point, we could not agree more.

References

Baker, M. Over half of psychology studies fail reproducibility test. *Nature* (2015). https://doi.org/10.1038/nature.2015.18248

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*, 2607–2612

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*, 435.

Button K. S. et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145–153. Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, *45*, 1304-1312.

Etz A, Vandekerckhove J (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLoS ONE 11(2)*: e0149794. https://doi.org/10.1371/journal.pone.0149794

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*(4), 576–588.

Fraley, R. C. & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. PLoS ONE 9(10): e109019. https://doi.org/10.1371/journal.pone.0109019

Gilbert DT, King G, Pettigrew S, Wilson TD. (2016). Comment on estimating the reproducibility of psychological science. *Science 351*(6277): 1037

Harris CR, Coburn N, Rohrer D, Pashler H (2013) Two Failures to Replicate High-Performance-Goal Priming Effects. PLoS ONE 8(8): e72467. https://doi.org/10.1371/journal.pone.0072467

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Ioannidis J. P. (2015). Failure to Replicate: Sound the Alarm. *Cerebrum: the Dana forum on brain science*, *2015*, cer-12a-15.

Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van 't Veer, A. E., Vranka, M., & Lakens, D. (2021). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000438

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524–532.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., …Nosek, B. L. (2018). Many labs 2: Investigating variation in replicability across

samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490.

Lewandowsky, S. & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*, 358 doi:10.1038/s41467-019-14203-0

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Malich, L., & Rehmann-Sutter, C. (2022). Metascience is not enough—A plea for psychological humanities in the wake of the replication crisis. *Review of General Psychology, 26*(2), 261–273.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600-2606.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* **349**:aac4716

Owens, B. (2018, November 19). Replication failures in psychology not due to differences in study populations. *Nature*. https://doi.org/10.1038/d41586-018-07474-y

Pashler H, Coburn N, Harris CR (2012) Priming of Social Distance? Failure to Replicate Effects on Social and Food Judgments. PLoS ONE 7(8): e42510. https://doi.org/10.1371/journal.pone.0042510

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Renkewitz, F., & Heene, M. (2019). The replication crisis and open science in psychology: Methodological challenges and developments [Editorial]. *Zeitschrift für Psychologie, 227*(4), 233–236.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE, 7*(3), Article e33423. https://doi.org/10.1371/journal.pone.0033423

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General, 144*(4), e73–e85.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant & Child Development*, **31**, e2295.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), Article 25152459211007467.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*(2), 309–316.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Szucs, D. & Ioannidis, J. P. A. (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol 15(3)*: e2000797. https://doi.org/10.1371/journal.pbio.2000797

van Assen, M. A., van Aert, R. C., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PloS one*, *9*(1), e84896.

Wilson, B. M., Harris, C. R. & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, *117*, 5559-5567.

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2022). Theoretical false positive psychology. *Psychonomic Bulletin & Review*, *29*(5), 1751-1775.

Wilson, B. M. & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*, 186-197.