In Press, Perspectives on Psychological Science, Sept 2012

Is the Replicability Crisis Overblown? Three Arguments Examined

Harold Pashler & Christine R. Harris

University of California, San Diego

ABSTRACT - ***We discuss three arguments voiced by scientists who view the current outpouring of concern about replicability as overblown. The first idea is that the adoption of a low alpha level (e.g., 5%) puts reasonable bounds on the rate at which errors can enter the published literature, making false-positive effects rare enough to be considered a minor issue. This, we point out, rests on statistical misunderstanding: The alpha level imposes no limit on the rate at which errors may arise in the literature (Ioannidis, 2005b). Second, some argue that whereas direct replication attempts are uncommon, conceptual replication attempts are common--providing an even better test of the validity of a phenomenon. We contend that performing conceptual rather than direct replication attempts interacts insidiously with publication bias, making it possible for literatures to emerge that appear to confirm the reality of phenomena that in fact do not exist. Finally, we discuss the argument that errors will eventually be pruned out of the literature if the field would just show a bit of patience. We contend that there are no plausible concrete scenarios to back up such forecasts, and that what is needed is not patience but rather systematic reforms in scientific practice.***

Address correspondence to Harold Pashler, Department of Psychology 0109, University of California, San Diego, La Jolla, CA 92093; e-mail: hpashler@ucsd.edu.

Whenever people speak of a "crisis" in any enterprise that has been around for a very long time--like Experimental Psychology (or Science in general)--a measure of skepticism is probably a very sensible reaction. Is the present flurry of concern about replicability and replication--the development that prompted the current special issue of *Perspectives on Psychological* Science--overblown? In this article, we explore the three arguments that we have heard most often from scientists, who see the current outpouring of concern over replicability as greatly overblown. These scientists, some of whom are prominent and accomplished researchers, view efforts to change scientific practices as unnecessary. We contend that these arguments are misguided in instructive ways. The first argument focuses on the rate of false positives (a topic afficianados of statistics and methodology are well familiar with)..

Argument 1: It is a given that there will be some nonzero rate of false positives, but scientists keep this tolerably low by setting a relatively conservative alpha level (e.g., 5%).

Thanks to the work of Ioannidis (2005b) and other statisticians, the essential problems with this view are already familiar to many. Nonetheless, it is our sense that a fairly large number of scientists in psychology and other fields are not familiar with this work and still tend (erroneously) to assume that alpha levels

represent upper bounds on the rate at which errors can accumulate in a literature.

So what is the truth of the matter? To put it simply, adopting an alpha level of, say, 5% means that about 5% of the time when researchers test a null hypothesis that is true--i.e., when they look for a difference that does not exist--they will end up with a statistically significant difference (a type 1 error or false positive[1].) Whereas some have argued that 5% would be too many mistakes to tolerate, it certainly would not constitute a flood of error. So what is the problem?

Unfortunately, the problem is that the alpha level does not provide even a rough estimate, much less a true upper bound, for the likelihood that any given positive finding appearing in a scientific literature will be erroneous. To estimate what the literature-wide false positive likelihood is, several additional values need to be specified, which can only be guessed at. We begin by considering some highly simplified scenarios. Although artificial, these have enough plausibility to provide some eye-opening conclusions.

For the following example, let us suppose that 10% of the effects that researchers look for actually exist, which will be referred to here as the prior probability of an effect (i.e., the null hypothesis is true 90% of the time). Given an alpha of 5%, Type 1 errors will occur in 4.5% of the studies performed (90% X 5%). If one assumes that studies all have a power of, say, 80% to detect those effects that do exist, correct rejections of the null hypothesis will occur in 8% of the time (80% X 10%). If one further imagines that all positive results are published then this would mean that the probability any given published positive result is erroneous would be equal to the proportion of false positives divided by the sum of the proportion of false positives plus the proportion of correct rejections. Given the proportions specified above, then, we see that more than one-third of published positive findings would be false positives (4.5% / (4.5% + 8%) = 36%). In this example, the errors occur at a rate approximately seven times the nominal alpha level (row 1 of Table 1).

INSERT TABLE 1 HERE

Table 1 shows a few more hypothetical examples of how the frequency of false positives in the literature would depend upon the assumed probability of null hypothesis being false and the statistical power. An 80% power likely exceeds any realistic assumptions about psychology studies in general. For example, Bakker, van Dijk, & Wikkerts, this issue, estimate .35 as a typical power level in the psychological literature. If one modifies the previous example to assume a more plausible power level of 35%, the likelihood of positive results being false rises to 56% (second row of the table). John Ioannides (2005) did pioneering work to analyze (much more carefully and realistically than we do here) the proportion of results that are likely to be false, and concluded that it could very easily be a majority of all reported effects.

For the probability that any given positive result in the published literature is wrong to occur as infrequently as the 5% alpha level, assuming 35% power, one would have to assume that differences looked for exist about 75% of the time (fourth row in the table.)

So, what is a reasonable estimate for the prior probability of effects that are tested by psychologists? Ioannidis (2005b) considers many domains in which the probability seems quite certain to be extremely low, e.g., epidemiological studies examining very long lists of dietary factors and relating them to cancer, or genome-wide association studies (GWAS) examining tens of thousands of genetic results. Experimental psychologists may think, "Well,

[1] Here we follow the standard approach of null hypothesis testing statistics and imagine that effects either exist or do not exist, ignoring the idea that differences of exactly zero may scarcely ever exist (see e.g., Nunnally, 1960). For those who find the arguments of Nunnally and others persuasive (as we do), it may be best to think of our discussions of "no effect" as meaning "no effect big enough to have any scientific interest".

that may be the case for massive exploratory studies in biomedicine, but I typically have some theoretically motivated predictions for which I have pretty high confidence". Thus, the reader may want to argue that a prior of 75% is not necessarily too high. However, in considering the likely credibility of the psychological literature as a whole, the issue is not whether investigators sometimes test hypotheses that they view as having a reasonable likelihood of yielding a positive result. The issue, rather, is the *totality* of effects that are tested and for which, given a positive result, investigators would proceed to publish the result and devise some theoretical interpretation. We suspect that if readers reflect on this question, they will conclude as we have done that this number is often quite large even in research that is, in a broad sense, theoretically motivated. As Kerr (1998) and Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) describe, experimental research is normally at least partly exploratory in nature even when it is presented in a confirmatory template. Thus, we would argue that the second row of the table probably comes closer to the situation in experimental psychology than we might like to imagine--implying that Ioannidis devastating surmise that "Most published results are false" could very easily be the case throughout our field. (Of course, this likelihood is amplified if journals are willing to publish surprising results based on a single positive finding--and we would argue that such a willingness is quite common, although certainly not universal.)

Thus, the fact that psychology and similar fields usually insist upon a reassuringly low alpha level (typically 5%) does not by any means imply that no more than 5% of the positive findings in the published literature are likely to be errors. Moreover, the situation is surely much worse than what the discussion above would suggest because in addition to testing many hypotheses with a low likelihood of effects, investigators often exploit hidden flexibility in their data analysis strategies, allowing the true alpha level to rise well above the nominal alpha level (Simmons et al, 2011; see also discussion by Ioannidis, 2005b, of "bias" and Wagenmakers et al., this issue, on "fairy tale factors".) Moreover, the highest impact journals famously tend to favor highly surprising results; this makes it easy to see how the proportion of false positive findings could be even higher in such journals than it would be in less career-enhancing outlets (Ioannidis, 2005a; Munafo et al., 2009). Naturally, those areas within psychology which lend themselves to performing a great number of tests on a variety of variables in any given study, as well as areas in which underpowered studies are more common, are likely more prone to false findings than are other areas.

In summary, our standard statistical practices provide no assurance that erroneous findings will occur in the literature at rates even close to the nominal alpha level (see Wagenmakers, Wetzels, Borsboom, van der Maas and Kievit, this issue, on the remarkably low diagnosticity of results meeting the standard criteria for rejecting the null hypothesis.) Given that errors are sure to be published in many cases, it seems to us that the most critical question is what happens to these errors after they are published. If they are often corrected, then the initial publication may not cause much harm.

Unfortunately, as many of the articles in the current special issue document and discuss (e.g., Makel et al., this issue), the sort of direct replications optimal for identifying erroneous findings are disturbingly rare. Moreover, even when such data are collected, the results are hard to publish, regardless of whether they confirm or disconfirm the finding. This brings us then to the second, and probably the most popular, response from defenders of the status quo.

Argument 2: It is true that researchers in many areas of psychology carry out direct replication attempts only rarely. However, researchers frequently attempt (and publish) conceptual replications, which are more effective than direct replications for assessing the reality and

importance of findings because they test not only the validity but also the generality of the finding.

This view was advocated, for example, by senior psychologists quoted by Carpenter (2012). In our opinion, this is a seductive but profoundly misleading argument. We contend that in any field where it is rare for people to conduct direct replications, and common to undertake conceptual replications, the field can be grossly misled about the reality of phenomena; and misled *even more gravely than would happen based on the issues discussed above*. The reason, it seems to us, is that conceptual replication attempts (especially when such studies are numerous and low in statistical power) interact in an insidious fashion with publication bias and also with the natural tendency for results perceived as "interesting" to circulate among scientists through informal channels.

To shed light on this, consider the question: When investigators undertake *direct* replications and they fail to obtain an effect, what are they likely to do with their results? In the ideal world, of course, they would publish these outcomes in journals--or make them public through other mechanisms such as websites or scholarly meetings. In fact, published non-replications are rare (e.g., Sterling, Rosenbaum & Weinkam, 1995; Makel et al., this issue). Nonetheless, we conjecture, when a respected investigator obtains but does not publish a negative result, the fact of the failure often achieves some limited degree of dissemination through informal channels. A failure to confirm a result based on a serious direct replication attempt is interesting gossip, and the fact is likely to circulate at least among a narrow group of interested parties. At a minimum, the investigator and his or her immediate colleagues will have reduced confidence in the effect.

By contrast, consider what happens when a scientific community undertakes only *conceptual* replication attempts. If a conceptual replication attempt fails, what happens next? Rarely, it seems to us, would the investigators themselves believe they have learned much of anything. We conjecture that the typical response of an investigator in this (not uncommon) situation is to think something like "I should have tried an experiment closer to the original procedure--my mistake." Whereas the investigator may conclude that the underlying effect is not as robust or generalizable as had been hoped, he or she is not likely to question the veracity of the original report. As with direct replication failures, the likelihood of being able to publish a conceptual replication failure in a journal is very low. But here, the failure will likely generate no gossip--there is nothing interesting enough to talk about here. The upshot, then, is that *a great many failures of conceptual replication attempts can take place without triggering any general skepticism of the phenomenon at issue* (see also Nosek, this issue, for a similar point.)

On the other hand, what happens when investigators undertake a conceptual replication and succeed? Without a doubt, such a success will be seen as interesting, and the researchers will seek to publish it, present it at meetings, etc. When they do so, they will likely receive encouragement from the original investigators (who are plausible reviewers for the work). Precisely because the conceptual replication attempt differs from the original study in procedure, it will often be seen as novel enough to warrant publication. In short, a conceptual replication success is much *more* publishable than would be a direct replication success--which is of course precisely why investigators are tempted to skip the direct replications and focus their efforts on conceptual replications.

The inevitable conclusion, it seems to us, is that a scientific culture and an incentive scheme that promotes and rewards conceptual rather than direct replications amplifies the publication bias problem in a rather insidious fashion. Such a scheme currently exists in every area of research of which we are aware.

We would go further and speculate that in any field where direct replication attempts are nonexistent and conceptual replication attempts are common, entire communities of researchers can easily be led to believe beyond question in phenomena that simply do not exist. What conditions are needed to promote such pathological results? The key element would seem to be that a pseudo-result appears both exciting and easy--the kind of result that would tempt hundreds of researchers to undertake small low-powered conceptual replication attempts (see Bakker, van Dijk, and Wicherts, this issue; Ioannidis, 2005b). Enough of these will "work" just by chance to generate the strong impression in the community that successful confirmations are plentiful (and if investigators exploit the hidden degrees of freedom pointed out by Simmons et al., 2011, there will be more of these than one would expect from the nominal alpha level.)

Pathological Science

We speculate that the harmful interaction of publication bias and a focus on conceptual rather than direct replications may even shed light on some of the famous and puzzling "pathological science" cases that embarrassed the natural sciences at several points in the 20th century, e.g., the Polywater (Rousseau & Porto, 1970) and Cold Fusion (Taubes, 1993) controversies. What many observers found peculiar in these cases was that it took many years for a complete consensus to emerge that the phenomena lacked any reality (and in the view of a few physicists, some degree of uncertainty may persist even to this day over Cold Fusion, although most physicists appear to regard the matter as having been settled decisively and negatively; Taubes, 1993.) Indeed, it appears that many exact replication attempts of the initial studies of Pons and Fleischman (who first claimed to have observed Cold Fusion) were undertaken soon after the first dramatic reports of cold fusion. Such attempts produced generally negative results (Taube, 1993). However, what kept faith in Cold Fusion alive for some time (at least in the eyes of some onlookers) was a trickle of positive results achieved using *very different designs* than the originals--i.e., what psychologists would call "conceptual replications".

This suggests that one important hint that a controversial finding is pathological may arise when defenders of a controversial effect disavow the initial methods used to obtain an effect and rest their case entirely upon later studies conducted using other methods. Of course, productive research into real phenomena often yields more refined and better ways of producing effects. But what should inspire doubt is any situation where a phenomenon, as presented by its defenders, becomes a "moving target" in terms of where and how it is elicited (cf Langmuir, 1953/1989). When this happens, it would seem sensible to ask: If the finding is real and yet the methods used by the original investigators are not reproducible, then how were these investigators able to uncover a valid phenomenon with methods that do not work? Again, the unavoidable conclusion is that a sound assessment of a controversial phenomenon should focus first and foremost on direct replications of the original reports, not on novel variations--each of which introduces independent ambiguities of its own.

This brings us to the third defense of the status quo to be discussed here – whether science is self-correcting in the long-term, even if it is error-prone in the short term.

Argument 3: Science is self-correcting but slow; whereas some erroneous results may get published, eventually these will be discarded. Current discussions of a replicability crisis reflect an unreasonable impatience.

This "long-term" optimism, which we have heard expressed quite frequently, seems to boil down to two very distinct arguments. The first is that if one just waits long enough, erroneous findings will actually be debunked in

an explicit fashion. Is there evidence that this sort of slow correction process is actually happening? Using Google Scholar we searched <"failure to replicate", psychology> and checked the first 40 articles among the search returns that reported a non-replication. Four years was the median time between the original target article and the replication attempt with only 10% of the replication attempts occurring at lags longer than 10 years (n=4). This suggests that when replication efforts are made (which, as already discussed, happens infrequently), they generally target very recent research. We see no sign that long-lag corrections are taking place.

A second version of this optimistic argument would contend that even if erroneous findings are rarely explicitly tested and refuted after substantial time delays, correction occurs in a different way. This view suggests that the long-term self-corrective process unfolds by field collectively "moving on" (in some nonrandom fashion) to focus on other (presumably more valid) phenomena. On this account, when the herd moves on it is a sign that better grazing land has been identified elsewhere. This "smart herd" metaphor strikes us as appealing but generally misleading.

Academic research is notoriously faddish, with nests of active researchers probing particular topics in one stretch of time, and other topics a few years later. The notion that when an area ceases to be an active focus of investigation, its research findings can and should be regarded as suspect seems bizarre if one thinks through its implications. Although from time to time, individual investigators undoubtedly do give up on a topic when they find that results do not replicate, that is only one of many reasons why research interests change. Others include: the questions have been satisfactorily answered, new techniques make other topics more appetizing, or tastes simply change for other reasons. Merely noting that active research on a topic has diminished does not separate out any of these cases.

To pick out just a few of a very large number of potential examples, in cognitive psychology there were waves of research at various times on such phenomena as selective attention to multi-channel speech stimuli, effects of imagery on long-term memory, and articulatory working memory. In each of these areas, there was notable progress (including quite a bit of direct replication of prior results), and none of these areas were (as far as we know) seen by experts as especially subject to replicability problems. Yet relatively fewer behavioral investigations of these topics appear to have been published in recent years. The hiatuses, we think, are a joint sign of the success of the work itself and the fact that interests simply moved on. Other highly specific factors may play a role: in the case of attention research, for example, the dramatic shift toward using visual rather than auditory stimuli was probably due to perceived methodological advantages (i.e., better temporal control). To assume that such shifts of research interest invalidate older bodies of empirical findings would, in our view, be ill-informed.

The notion that declines in research activity on a given topic indicates that the empirical literature in that area is likely invalid also flies in the face of the practices of those writing textbooks and review articles. For example, the bodies of research mentioned in the preceding paragraph continue to be discussed in leading textbooks (e.g., Reisberg, 2009), and we see no reason to think that they do not deserve such citation. Indeed, an assumption that older literatures are likely to be invalid would make a mockery of meta-analytic efforts, such as those seeking to uncover the causes of diseases by comparing effect sizes for predictor variables studied in widely scattered literatures over many decades--efforts that have received highly positive reviews in the field (see, e.g., Heinrichs, 2001, on schizophrenia). Whereas older as well as recent bodies of work undoubtedly contain errors, there is no reason to believe that research

fads provide a valid indicator of the solidity of different topics.

The non-self-correcting nature of science has been highlighted lately in the pharmaceutical domain, illustrating the serious consequences of current practices. Recent reports emerging from the US pharmaceutical industry reveal the extent to which an accumulation of errors in the basic research literature can obstruct translational progress. Writing in *Nature*, Begley and Ellis (2012) described the first-hand experience of the Amgen Corporation over a 10-year period in attempting to build drug development programs upon 53 "landmark" published studies in preclinical (basic) cancer research. Despite systematic and strenuous efforts, they found that only six of the phenomena they examined (11%) could be replicated. Interestingly, they noted that, "Some non-reproducible preclinical papers had spawned an entire field, with hundreds of secondary publications that expanded on elements of the original observation, but did not actually seek to confirm or falsify its fundamental basis." A scientist from another major pharmaceutical company told a reporter "It drives people in industry crazy. Why are we seeing a collapse of the pharma and biotech industries? One possibility is that academia is not providing accurate findings." (CNBC, 2012). These discussions show that invalid basic research findings frustrate the long-term translational process, and there is no reason to suppose that the frustration itself feeds back to correct errors in the basic-science literature (for example, in the case described by Begley and Ellis, 2012, it does not appear that Amgen's failures to replicate were ever published, so the erroneous results may continue to misdirect drug development efforts for years to come.)

In summary, we would argue that it appears almost certain that fallacious results are entering the literature at worrisome rates. More precise information about the rate at which this is happening in psychology should begin to emerge in due course from the Replicability Project described by Nosek (this issue). Unfortunately, however, there is every reason to believe that of the errors that do enter the literature, the great majority will persist uncorrected essentially indefinitely, given current practices. Errors will be propagated through textbooks and review articles and people interested in a topic will be misinformed for generations. Finally, as the experiences of Begley and Ellis (2012) suggest, the long-term harm may not be limited to confusion; errors may also stymie the development of practical applications from basic research. At the very least, then, it seems to us that the onus is on anyone defending the status quo to articulate exactly how the delayed self-correction process they envision is supposed to operate--and to show examples of where it is working effectively.

Concluding Remarks

In closing, we have considered here three arguments offered by those who view current concerns about the rate of replicability problems in psychology as overblown. We have contended that these arguments do not comport with the readily observable practices and habits of investigators in the behavioral sciences. In our view, there are likely to be serious replicability problems in our field, and correcting these errors will require many significant reforms in current practices and incentives. The possible directions for such reforms are discussed in many of the articles in the current issue.

Acknowledgements

The authors are grateful to Jon Baron, Vic Ferreira, Alex Holcombe, Dave Huber, Keith Rayner, Tim Rickard, Roddy Roediger, Ed Vul, and John Wixted for useful comments and discussion, and Noriko Coburn for assistance in preparation of the manuscript.

REFERENCES

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science,* this issue.

Begley, C. G., and Ellis, L. M. Improve standards for preclinical cancer research. *Nature,* 2012 Mar 28;483(7391):531-3. doi: 10.1038/483531a.

Carpenter, S. (2012). Psychology's bold initiative: In an unusual attempt at scientific self-examination, psychology researchers are scrutinizing their field's reproducibility. *Science, Vol. 335,* 30 March 2012, 1558-1560.

CNBC (2012). Many Cancer Studies Produce Unreliable Results: Study. http://www.cnbc.com/id/46882434

Csada RD, James PC, Espie RHM (1996) The "file drawer problem" of non-significant results: Does it apply to biological research? *Oikos* 76: 591–593.

Heinrichs, S. (2001). *In Search of Madness: Schizophrenia and Neuroscience.* Oxford, UK: Oxford University Press.

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *J. Am. Med. Assoc.* 294, 218–228.

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine,* 2, 696–701.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217.

Langmuir, I. (1953/1989). Pathological science (Transcript of a talk given at The Knolls Research Laboratory, Schenectady, New York, December 18, 1953.) *Physics Today, 42,* 36–48.

Munafò, M. R., Stothart, G., and Flint, J. (2009). Bias in genetic association studies and impact factor. *Molecular Psychiatry,* 14, 119–120.

Neuliep, J., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior & Personality,* 8, 21-29.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement, 20,* 641–650.

Reisberg, D. (2009). *Cognition, Exploring the Science of the Mind, 4th Edition*. Norton.

Rousseau, D. L., & Porto, S. P. S. (1970). Polywater: polymer or artifact. *Science, Vol. 167, no. 3926,* 1715-1719.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science,* 22, 1359–1366.

Sterling, T. D., Rosenbaum, W. L., and Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician,* 439: 108–112.

Taubes, G. (1993). *Bad science: The short life and weird times of cold fusion.* Random House.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100,* 426–432.

Prior Probability of Effect	Power	Proportion of Studies Yielding True Positives	Proportion of Studies Yielding False Positives	Proportion of Positive Results that are False
10%	80%	8%	4.5%	36%
10%	35%	3.5%	4.5%	56%
50%	35%	17.5%	2.5%	13%
75%	35%	26.3%	1.6%	5%

Table 1: Proportion of positive results that are false given assumptions about prior probability of an effect and power.