aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 645–654

DOI: 10.1177/1745691612464056
http://pps.sagepub.com
SAGE

# Why Science Is Not Necessarily Self-Correcting

**John P. A. Ioannidis**
Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences

## Abstract

The ability to self-correct is considered a hallmark of science. However, self-correction does not always happen to scientific evidence by default. The trajectory of scientific credibility can fluctuate over time, both for defined scientific fields and for science at-large. History suggests that major catastrophes in scientific credibility are unfortunately possible and the argument that "it is obvious that progress is made" is weak. Careful evaluation of the current status of credibility of various scientific fields is important in order to understand any credibility deficits and how one could obtain and establish more trustworthy results. Efficient and unbiased replication mechanisms are essential for maintaining high levels of scientific credibility. Depending on the types of results obtained in the discovery and replication phases, there are different paradigms of research: optimal, self-correcting, false nonreplication, and perpetuated fallacy. In the absence of replication efforts, one is left with unconfirmed (genuine) discoveries and unchallenged fallacies. In several fields of investigation, including many areas of psychological science, perpetuated and unchallenged fallacies may comprise the majority of the circulating evidence. I catalogue a number of impediments to self-correction that have been empirically studied in psychological science. Finally, I discuss some proposed solutions to promote sound replication practices enhancing the credibility of scientific results as well as some potential disadvantages of each of them. Any deviation from the principle that seeking the truth has priority over any other goals may be seriously damaging to the self-correcting functions of science.

## Keywords

self-correction, replication

## Scientific Credibility: A Fluctuating Trajectory

Self-correction is considered a key hallmark of science (Merton, 1942, 1973). Science, by default, does not adopt any unquestionable dogma—all empirical observations and results are subject to verification and thus may be shown to be correct or wrong. Sooner or later, if something is wrong, a replication effort will show it to be wrong and the scientific record will be corrected.

The self-correction principle does not mean that all science is correct and credible. A more interesting issue than this eschatological promise is to understand what proportion of scientific findings are correct (i.e., the credibility of available scientific results). One can focus on the credibility of new, first-proposed results or on the credibility of all available scientific evidence at any time point. The proportion of first-proposed results (new discoveries, new research findings) at any time point (say, those new discoveries first published in 2012) that are correct can be anywhere from 0% to 100% (Ioannidis, 2005). The proportion of correct results, when all available data until that time-point are integrated (e.g., the latest updates of cumulative meta-analyses considering all the available evidence), can also vary from 0% to 100%. If we consider two different years, T1 and a later T2, the credibility of new proposed research findings from T2 may be larger, smaller, or the same as the credibility of new proposed research findings from T1. Similarly, the credibility of the updated evidence from T2 may be larger, smaller, or the same as the credibility of the evidence from T1.

Even if we believe that properly conducted science will asymptotically trend towards perfect credibility, there is no guarantee that scientific credibility continuously improves and that there are no gap periods during which scientific credibility drops or sinks (slightly or dramatically). The credibility of new findings and the total evidence is in continuous flux. It may get better or worse. And, hopefully, there should be some

**Corresponding Author:**
John P.A. Ioannidis, Stanford Prevention Research Center, 1265 Welch Rd, MSOB X306, Stanford University School of Medicine, Stanford, CA 94305
E-mail: jioannid@stanford.edu

ways that allow us to help make it better. The study of the trajectory of the credibility of scientific findings and of ways to improve it is an important scientific field on its own.

One may argue that, under an optimistic scenario, if the time period between T1 and T2 is large enough, then the scientific credibility will likely improve. However, we have inadequate evidence on how long that period has to be so as to have 90%, 95%, or 99% chances of seeing an improvement. Moreover, even when we do witness an overall improvement, it is likely that there will be substantial heterogeneity across scientific fields: Credibility may increase in some fields, but may decrease in others. If fields with low credibility become prolific in their productivity and dominant (e.g., they attract many researchers, funding, and editorial and public attention), then they may decrease the overall credibility of the scientific corpus, even if major progress is made in many other fields.

## Learning From History

Most short-term fluctuations in the credibility of scientific results over time are likely to be modest improvements or modest deteriorations. However, historical lessons suggest that major drops in credibility in large fields or even across the scientific continuum are possible. Major drops in credibility can occur by massive destruction of evidence or massive production of wrong evidence or distortion of evidence.

The fate of the Library of Alexandria offers one example of multiple massive destructions. The largest library of the ancient world, containing probably over 1 million items at its peak, was destroyed at least 4 times (twice by Roman wars, once by Christian mobs, and finally after the Arab conquest) until its total extinction.

Phrenology is one example of massive production of wrong evidence. It was conceived by a physician, Franz Joseph Gall, in 1796 and dominated neuropsychology in the 19th century. I copy from Wikipedia (http://en.wikipedia.org/wiki/Phrenology, accessed July 2012):

> Phrenology was a complex process that involved feeling the bumps in the skull to determine an individual's psychological attributes. Franz Joseph Gall first believed that the brain was made up of 27 individual "organs" that determined personality, with the first 19 of these "organs" believed to exist in other animal species. Phrenologists would run their fingertips and palms over the skulls of their patients to feel for enlargements or indentations. The phrenologist would usually take measurements of the overall head size using a caliper. With this information, the phrenologist would assess the character and temperament of the patient and address each of the 27 "brain organs." Gall's list of the "brain organs" was lengthy and specific, as he believed that each bump or indentation in a patient's skull corresponded to his "brain map." An enlarged bump meant that the patient utilized that particular "organ" extensively. The 27 areas were varied in function, from sense of color, to the likelihood of religiosity, to the potential to commit murder.

There are many other historical examples of massive production of wrong evidence or distortion of evidence: the distortion of eugenics in Nazi Germany so as to show the superiority of the Arians (Aly, 1994), and the massively-funded research by the tobacco industry in the 20th century supposedly to evaluate the "unclear" consequences of smoking (Proctor, 2011).

In settings where science is at a low ebb and massive destruction of evidence, production of wrong evidence, or distortion of evidence abounds, it is possible that the scientific environment becomes so perverted that people don't even perceive that this is happening and thus they do not worry about it. They feel that their current science is robust and most correct. The Christian mobs destroying the Library of Alexandria, phrenologists, and Nazi eugenicists must have felt quite secure and self-justified in their dogmas.

## Current Fluctuations in the Credibility of Science

How is current science faring? Obviously, we see tremendous achievements in technology, measurement ability, and in the amount and sometimes even the quality of data. However, the state of the evidence needs careful appraisal. It could be that the credibility of some disciplines is improving, whereas some others may be facing difficult times with decreased credibility. It is not possible to exclude even the possibility that massive drops in credibility could happen or are happening.

For example, is it possible that we are facing a situation where there is massive destruction of evidence? At first sight, this would sound weird, as current science is apparently so tolerant. Obviously, book burning is a despicable, unacceptable behavior according to current norms. However, it is also possible that a Library of Alexandria actually disappears every few minutes. Currently, there are petabytes of scientific information produced on a daily basis and millions of papers are being published annually. In most scientific fields, the vast majority of the collected data, protocols, and analyses are not available and/or disappear soon after or even before publication. If one tries to identify the raw data and protocols of papers published only 20 years ago, it is likely that very little is currently available. Even for papers published this week, readily available raw data, protocols, and analysis codes would be the exception rather than the rule. The large majority of currently published papers are mostly synoptic advertisements of the actual research. One cannot even try to reproduce the results based on what is available in the published word.

Moreover, is it possible that we are currently facing a situation where there is massive production of wrong information or distortion of information? For example, could it be that the advent of research fields in which the prime motive and strongest focus is making new discoveries and chasing statistical

significance at all cost has eroded the credibility of science and credibility is decreasing over time?

Empirical evidence from diverse fields suggests that when efforts are made to repeat or reproduce published research, the repeatability and reproducibility is dismal (Begley & Ellis, 2012; Donoho, Maleki, Rahman, Shahram, & Stodden, 2009; Hothorn & Leisch, 2011; Ioannidis et al., 2009; Prinz, Schlange, & Asadullah, 2011). Not surprisingly, even hedge funds don't put much trust on published scientific results (Osherovich, 2011).

## Science at-Large on Planet F345, Andromeda Galaxy, Year 3045268

Planet F345 in the Andromeda galaxy is inhabited by a highly intelligent humanoid species very similar to *Homo sapiens sapiens*. Here is the situation of science in the year 3045268 in that planet. Although there is considerable growth and diversity of scientific fields, the lion's share of the research enterprise is conducted in a relatively limited number of very popular fields, each one of that attracting the efforts of tens of thousands of investigators and including hundreds of thousands of papers. Based on what we know from other civilizations in other galaxies, the majority of these fields are null fields—that is, fields where empirically it has been shown that there are very few or even no genuine nonnull effects to be discovered, thus whatever claims for discovery are made are mostly just the result of random error, bias, or both. The produced discoveries are just estimating the net bias operating in each of these null fields. Examples of such null fields are nutribogus epidemiology, pompompomics, social psychojunkology, and all the multifarious disciplines of brown cockroach research—brown cockroaches are considered to provide adequate models that can be readily extended to humanoids. Unfortunately, F345 scientists do not know that these are null fields and don't even suspect that they are wasting their effort and their lives in these scientific bubbles.

Young investigators are taught early on that the only thing that matters is making new discoveries and finding statistically significant results at all cost. In a typical research team at any prestigious university in F345, dozens of pre-docs and post-docs sit day and night in front of their powerful computers in a common hall perpetually data dredging through huge databases. Whoever gets an extraordinary enough omega value (a number derived from some sort of statistical selection process) runs to the office of the senior investigator and proposes to write and submit a manuscript. The senior investigator gets all these glaring results and then allows only the manuscripts with the most extravagant results to move forward. The most prestigious journals do the same. Funding agencies do the same. Universities are practically run by financial officers that know nothing about science (and couldn't care less about it), but are strong at maximizing financial gains. University presidents, provosts, and deans are mostly puppets good enough only for commencement speeches and other boring ceremonies and for making enthusiastic statements about new discoveries of that sort made at their institutions. Most of the financial officers of research institutions are recruited after successful careers as real estate agents, managers in supermarket chains, or employees in other corporate structures where they have proven that they can cut cost and make more money for their companies. Researchers advance if they make more extreme, extravagant claims and thus publish extravagant results, which get more funding even though almost all of them are wrong.

No one is interested in replicating anything in F345. Replication is considered a despicable exercise suitable only for idiots capable only of me-too mimicking, and it is definitely not serious science. The members of the royal and national academies of science are those who are most successful and prolific in the process of producing wrong results. Several types of research are conducted by industry, and in some fields such as clinical medicine this is almost always the case. The main motive is again to get extravagant results, so as to license new medical treatments, tests, and other technology and make more money, even though these treatments don't really work. Studies are designed in a way so as to make sure that they will produce results with good enough omega values or at least allow some manipulation to produce nice-looking omega values.

Simple citizens are bombarded from the mass media on a daily basis with announcements about new discoveries, although no serious discovery has been made in F345 for many years now. Critical thinking and questioning is generally discredited in most countries in F345. At some point, the free markets destroyed the countries with democratic constitutions and freedom of thought, because it was felt that free and critical thinking was a nuisance. As a result, for example, the highest salaries for scientists and the most sophisticated research infrastructure are to be found in totalitarian countries with lack of freedom of speech or huge social inequalities—one of the most common being gender inequalities against men (e.g., men cannot drive a car and when they appear in public their whole body, including their head, must be covered with a heavy pink cloth). Science is flourishing where free thinking and critical questioning are rigorously restricted, since free thinking and critical questioning (including of course efforts for replicating claimed discoveries) are considered anathema for good science in F345.

## But Progress is Made, No?

I don't even want to think that Earth in 2012 AD is a replica of F345 in year 3045268. However, there are some features where our current modes of performing, reporting, and replicating (or not replicating) scientific results could resemble this dreadful nightmare. More important, we may well evolve toward the F345 paradigm, unless we continuously safeguard scientific principles. Safeguarding scientific principles is not something to be done once and for all. It is a challenge that needs to be met successfully on a daily basis both by single scientists and the whole scientific establishment. Science may

well be the noblest achievement of human civilization, but this achievement is not irreversible.

Many investigators may dismiss the F345 scenario and may be very optimistic about our current state of scientific progress. A main argument is that, leaving theoretical concerns about credibility aside, the practical consequences of scientific progress are readily visible in our everyday lives. We have developed increasingly powerful computers, airplanes, space shuttles, and have extended life expectancy. Some progress is clearly made, so it *must be* that science is powerful and that we know a lot.

This argument is very weak. The fact that some practical progress is made does not mean that scientific progress is happening in an efficient way or that we cannot become even more efficient. I guess that some high priests of the Egyptians may have equally claimed that their science was perfect and optimal because they had discovered fire and wheels—what more could one hope to achieve? They may have laughed at someone who might have claimed that we could find anything more sophisticated than fire and wheels.

Furthermore, even if we have an increasing number of successful, hopefully correct, scientific results (that also "work"), this says nothing about the proportion of scientific results that are not correct. The number of scientific results in some fields increases exponentially over time. If the number of correct results increases but the number of wrong results increases more rapidly, then the scientific credibility overall decreases.

Finally, arguments linking practical progress to credibility are making a logical leap of attribution. Even when some progress in human affairs is documented, it is not at all certain that we know where to attribute it. For example, extensions in life expectancy are modest (at best) in developed countries in recent years. It is not at all clear that the majority of improvements in life expectancy are due to medicine (medical science and health care) rather than other practical improvements (e.g., improvements in hygiene, sanitation, housing, communication, and overall level of living). Developed countries have spent over $10 trillion of resources annually on medicine (including all health care) in the last decade, and the pace is accelerating. If the return is very modest, there is even a chance that wrong and inefficient medicine is currently becoming a major disaster for humans and human civilization. Investing these resources elsewhere would have led to much greater benefits. Withholding these resources so as to invest them in medicine may have caused more lives to be lost prematurely. Medicine and healthcare may still be very successful in destroying human civilization in the near future if expenditures of cost-inefficient medicine based on no, limited, or flawed scientific evidence (or ignoring correct evidence) continue to escalate.

## Discovery and Replication

The current issue of *Perspectives on Psychological Science* contains a number of very interesting articles on the reproducibility of research findings, with emphasis on psychological science in particular. The contributed papers investigate how psychologists try to present perfect, significant results so as to survive in the profession through the publication bottleneck (Giner-Sorola, 2012, this issue); how often replication efforts are published in the psychological literature, what kind of replications these are (conceptual or direct), and who publishes them (Makel, Plucker, & Hegarty, 2012, this issue); why replication is not popular in current scientific circles and why there is demonstrably an excess of statistically significant results in the literature indicative of strong biases (Bakker, van Dijk, & Wicherts, 2012, this issue; Francis, 2012a, this issue); how the vast majority of analyses in psychological science are fine tuned to obtain a desired result (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012, this issue), but also counterarguments of how that bias may eventually become less of an issue (Galak & Meyvis, 2012, this issue); why the replication crisis is serious and how science won't necessarily correct itself unless direct (not just conceptual) replications are performed (Pashler & Harris, 2012, this issue); how one could change the focus of science from getting something novel and significant to getting to the truth (Nosek, Spies, & Motyl, 2012, this issue); what psychologists think about the current system and how they are open to changes, but also weary of too much emphasis on new rules that may have to be fulfilled (Fuchs, 2012, this issue); and how students may be a potential workforce who may perform some of the required replication work (Frank & Saxe, 2012, this issue; Grahe et al., 2012, this issue). Given this emerging picture, psychological science seems to be facing similar challenges to many other scientific domains that struggle through these problems and many scientists are at risk at habituating eventually the nine circles of scientific hell described by Neuroskeptic (2012, this issue). It would be useful to use the evidence on psychological science from these papers and from the previous literature that these papers summarize to understand some issues about the interface between discovery and replication, impediments to the self-correction of science, and whether proposed solutions for fixing these problems are likely to work or not.

## Possible Pairs of Discovery and Replication Results and Their Prevalence

Table 1 shows in a simplified way the possible pairs of discovery and replication results when a scientific finding is claimed, defined here as the discovery of some effect for which the null hypothesis ($H_0$) is rejected. Bakker et al. (2012) estimate that almost all the papers in psychological science claim "positive" findings anyhow. Of note, this ubiquity of "positive" results has been well-documented in very diverse scientific fields, and it seems to be increasing in disciplines in lower positions in the hierarchy of evidence (Fanelli, 2010a, 2010b, 2012).

As shown in Table 1, the optimal paradigm is to make a correct discovery and to correctly replicate it. The self-correcting paradigm occurs when the discovery is wrong, but replication allows us to correct the error. Both of these paradigms are eventually favorable for the credibility of scientific evidence.

**Table 1.** Possibilities of Discovery and Replication: Six Possible Paradigms

| | Replication results | | |
|---|---|---|---|
| Discovery results | Correct | Wrong | Not obtained |
| Correct (true positive) | Optimal: ≤ 1%* | False nonreplication: <<1%* | Unconfirmed genuine discovery: 43%** |
| Wrong (false positive) | Self-correcting: ≤1%* | Perpetuated fallacy: 2%* | Unchallenged fallacy: 53%** |

*The sum of the items in the first two columns is assumed to be close to 4%, a probably generous estimate vis-à-vis the data by Makel et al. (2012).

**Estimate of 53% for unchallenged fallacy is indicative and based on Pashler and Harris (2012) estimate of 56% for all false-positives combined (self-correcting, perpetuated fallacy, unchallenged fallacy) assuming a prior probability of a nonnull effect of 10%. However, depending on assumptions about prior odds of a genuine nonnull effect and whether bias is present or not, unchallenged fallacy may easily have a prevalence of 30% to 95%, and unconfirmed genuine discoveries may represent the remaining 66% to 1%, respectively. Different subfields in psychological science may have different prevalence of unconfirmed genuine discoveries and unchallenged fallacies within these ranges.

Problems occur with two other paradigms. False nonreplication occurs when the original discovery is correct, but the replication is wrong. Important discoveries may be unwisely discredited. Perpetuated fallacy occurs when both the discovery and the replication are wrong (e.g., because the same errors or biases (or different ones) distort the results). Finally, in the absence of any replication efforts, one is left with unconfirmed genuine discoveries and unchallenged fallacies.

This classification is perhaps oversimplified. One could also consider effect sizes and the distance of the observed effect size in discovery studies from the true effect size. Replication efforts could lead closer to or further from the truth. However, the issues would still be similar. The question is what the prevalence of each of these patterns is in the scientific literature.

Based on the papers presented in this issue of *Perspectives*, it seems that the total prevalence of the first four paradigms (those where replication has been attempted following "positive" findings) is very low, in the range of 1%–5%. On the basis of Makel et al.'s (2012) article, replications (at least published ones) in psychological science are extremely uncommon—in the same range as previously documented for economics, marketing, and communication (Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2006; Hubbard & Armstrong, 1994; Kelly, Vhase, & Tucker, 1979). Makel et al. (2012) screened all the papers published by the 100 most-cited psychology journals since 1900 and they estimated that only 1.07% of these papers represented replication efforts. Even if we allow that not all other papers claim discoveries (e.g., some articles may have no original data at all, but may be reviews, editorials, letters, or other viewpoints), that replication papers become more frequent in recent years, and that a simple search using "replicat*" may have missed a fraction of the published replication efforts, it is unlikely that replication efforts exceed 5% of the current psychological literature with original data and may well be closer to 1%. Of the 1.07% identified by Makel et al. (2012), only 18% included direct rather than just conceptual replications and only 47% were done by investigators entirely different than the original authors. Although we are not given the cross-tabulation of direct replications and authors, it seems that direct replications by different authors than those who proposed the original findings are exceptionally uncommon and only a fraction of those are successful.

The self-correcting paradigm also seems to be very uncommon, as clear refutations of proposed discoveries are even less common than "successful" replications. Perhaps, this is not surprising, since independent replication seems to have low status in the psychological community and not too many investigators venture to refute what they or others have proposed, even if they clearly see that these propositions are wrong. If there is no reward, or even embarrassment and penalties, from being a critical replicator, self-correction from independent efforts will not flourish. The Proteus phenomenon (Ioannidis & Trikalinos, 2005) has been described in some other scientific disciplines, where publication of refutations may be attractive, when the original discovery has drawn a lot of attention and when the refutation is easy and quick to obtain. Perhaps, most psychological studies are not as easy to perform as, say, genetic association studies; investigators thus ponder if they should waste their limited time and resources toward killing someone else's claims and getting only headaches instead of any credit.

On the basis of these same arguments, false nonreplication (a subset of the very uncommon nonreplications) must be very rare. However, it is a type of paradigm that deserves careful study. The correctness of refutation in a replication study cannot be taken for granted, and it requires careful scrutiny to understand why replication efforts dissent against original discoveries. Lack of sufficient power and bias in the replication efforts are two possible explanations.

It is possible that, in several fields in psychological science, the current dominant paradigm when replication is attempted is that of perpetuated fallacies. Replication efforts, rare as they are, are done primarily by the same investigators who propose the original discoveries. Many of these replications are simply conceptual replications (Makel et al., 2012; Nosek et al.,

2012), which probably suffer from confirmation bias (Wagenmakers et al., 2012) in which authors try to get the result they want (the "fine-tune factor," or what I have previously described as "vibration of effects" (Ioannidis, 2008). Some other replication efforts may be done by the same investigators in different papers or even by different investigators who nevertheless have strong allegiance bias (e.g. they strongly believe in the same theory and expect to get the same results, reinforcing the status of what was originally proposed).

The claimed discoveries that have no published replication attempts apparently make up the vast majority of psychological science. With an average power of 35% for psychological investigations estimated by Bakker et al. (2012), if the prior probability of a nonnull effect in the broad universe of all analyses undertaken by psychologists is 10%, Pashler & Harris, 2012 estimated that 56% of the original research findings may be false positives and, in the absence of any replication attempts, the vast majority remain unchallenged fallacies. This calculation actually does not consider the effect of potential biases (publication biases, other selection biases, etc.). If there is also modest bias (Ioannidis, 2005), then the prevalence of unchallenged fallacies may represent even up to 95% (if not more) of the significant findings in some areas of the psychological literature (Table 1). It is possible that different psychological science subfields have different priors and different biases, so it would not be surprising if the proportion of unchallenged fallacies varies considerably across subfields (e.g., from 30% to 95%). Then, the remaining 66% to 1%, respectively, would be unconfirmed genuine discoveries. In all, the overall credibility of psychological science at the moment may be in serious trouble.

## Impediments to Self-Correction: A View From Psychological Science

Table 2 summarizes some impediments to self-correction in science. Psychology was apparently the first scientific discipline to recognize the importance of publication bias (Rosenthal, 1979) and many of the methods that have been developed to try to probe for publication bias or adjust for its presence have also been developed in the psychological literature. Publication bias seems to be common (Ferguson & Brannick, 2012; Shadish, Doherty, & Montgomery, 1989), although by definition it is not possible to have a precise estimate in the absence of precise preregistration of studies. It is possible that other types of selective reporting bias surrounding the analyses and presentation of outcomes of scientific investigations may be more common than classic publication bias (Dwan et al., 2011). Classic publication bias considers that there are specific well-delineated studies with clear protocols, data, and analyses that disappear completely in a file drawer. In psychological science, as well as in other scientific fields, a study may be poorly defined and no protocol may exist. Investigators may continue adding and melding data, analyses, and subanalyses until something significant and publishable emerges. Several empirical studies in psychological science have highlighted some of the mechanisms that lead to such selective reporting without necessarily classic file-drawer type of publication bias (Bakker & Wicherts, 2011; Fiedler, 2011; Simmons, Nelson, & Simonsohn, 2011) and Wagenmakers et al. (2012) make a lively presentation of how this fine tuning may occur. Occasionally, results may even be entirely fabricated (Fanelli, 2009). The distinction between creative exploratory analysis, falsification, and fraud should ideally be easy to make, but in real life it is not always so. Frank fabrication where all the data do not exist at all is probably uncommon, but other *formes frustes* of fraud may not be uncommon, and questionable research practices are probably very common (John, Loewenstein, & Prelec, 2012).

Eventually all these aforementioned biases may converge towards generating an excess of significant results in the literature. Methods have been developed to test for excess of significance across sets of studies, meta-analyses, and entire fields (Ioannidis & Trikalinos, 2007). Variations thereof have been applied also in psychological science, based on the

**Table 2.** A List of Described Impediments to Self-Correction in Science With Reference to Psychological Science

| Impediments | Selected references |
|---|---|
| Publication bias | Ferguson & Brannick (2012); Shadish et al. (1989) |
| Other selective reporting bias (analysis and outcomes) | |
| Flexibility in data collection and analysis | Simmons et al. (2011); Wagenmakers et al. (2012); |
| Misreporting of results | Bakker & Wicherts (2011) |
| Voodoo correlations | Fiedler (2011) |
| Fabricated results | Fanneli (2009) |
| Other questionable research practices | John et al. (2012) |
| Excess significance bias (may reflect any of the above) | Francis (2012b); Ioannidis & Trikalinos (2007) |
| Underpowered studies | Maxwell (2004) |
| No replication work done—especially direct replication by independent investigators | Makel et al. (2012) |
| Underestimation of the replication crisis | Pashler & Harris (2012) |
| Editorial bias against replication research | Neuliep & Crandall (1990) |
| Reviewer bias against replication research | Neuliep & Crandall (1993) |
| Data, analyses, protocols not publicly available | Alsheikh-Ali et al. (2011); Wicherts, Borsboom, Kats, & Molenaar(2006) |

premise of examining whether the results are too good to be true (Francis, 2012b). One should examine such empirical studies cautiously, as sometimes genuine heterogeneity when addressing different questions or different experiments, may masquerade as excess significance. However, it is striking that the proportion of significant results in psychological science is approaching 100%, and this has been noticed even half a century ago (Sterling, 1959) with no change in the interim (Bakker et al., 2012; Sterling, Rosenbaum, & Weinkam, 1995). This preponderance of nominally significant results is astonishing, given that psychological investigations remain largely underpowered (Maxwell, 2004). Moreover, what is not appreciated is that underpowered studies not only have low chances of detecting effects that are nonnull; they also have a lower positive predictive value for the results to be correct when they identify a statistically significant effect (Ioannidis, 2005), and they are likely to identify estimates of effects that are exaggerated, even when a genuinely nonnull effect is discovered (Ioannidis, 2008). Thus, even if some effects in psychological science are genuinely nonnull, their original estimates may be unrealistically large when they are first discovered.

For example, Galak and Meyvis (2012) rebuke the criticism of Francis on their study by admitting that they do have unpublished results in the file drawer but even if these unpublished results are considered, a meta-analysis of all the data would still show a highly statistically significant effect. It is interesting to note that they admit that the unpublished results do show substantially smaller effects than the published results, so the true effect is apparently smaller than what is visible in the published record. Galak and Meyvis claim that a meta-analysis where all authors are asked to contribute their unpublished file-drawers would be a simple way to arrive at unbiased effect estimates. However, this has not worked well in my experience in diverse fields where I have attempted to perform meta-analyses. First, it is notoriously difficult to get primary unpublished data from all authors who have published something already. Second, some authors may have only unpublished data and thus there would be no way to know of them (unless the studies have already been registered). Finally, one can never be certain that, if some additional data are provided, this is the entire file drawer or just a favorable enough part of it. For example, in medicine, some analyses including file-drawer data from the industry may be considered less reliable than those depending just on published data.

One would expect that false positives and exaggerated effect estimates will be corrected, if proper replication is considered. This will be demonstrable as a decline effect in updated meta-analyses. However, in the presence of publication and other selective reporting biases, this correction will not be seen necessarily (Schooler, 2011) and perpetuated fallacies may ensue. There has been good documentation that replication in psychological science is rarely performed (Makel et al., 2012) and is undervalued both by editors and by peer reviewers (Neuliep & Crandall, 1990, 1993). Pashler and Harris (2012) describe (and rebut) the common arguments that replication is not a big issue and the replication crisis is not serious. Finally, the lack of routine open access to raw data, analysis codes, and protocols (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011; Wicherts et al., 2006) does not allow outside investigators to repeat, and thus verify, the published results. This lack of access also acts as an impediment towards integrative analyses using raw, individual-level data.

There are some additional potential impediments that have not been discussed specifically in psychological science but may be worthwhile noting here. These include the lack of a tradition for large-scale consortium-type prospective research and less emphasis on standards for conduct and reporting of research. Multicenter studies are probably less common in psychological science than in many other life sciences, and large-scale consortia with tens and hundreds of collaborating teams (as has been the case in genomics, for example) are not commonplace. Moreover, in contrast to life sciences where guidelines on how to conduct experiments (e.g., microarray experiments) and reporting guidelines have been extensively developed and very popular (Simera, Moher, Hoey, Schulz, & Altman, 2010), this process has attracted yet less traction in psychological science.

Conversely, most of psychological science may not face some other impediments to self-correction that are common in other life sciences. Most notable is the lesser penetration of financial and corporate conflicts of interest. Such conflicts are probably common in the medical sciences, where healthcare carries a huge budget attracting the pharmaceutical, medical devices, and technologies industry. Professional corporate involvement does not cause the quality of research to deteriorate. In fact, research may even improve in terms of some widely accepted indicators of quality (e.g., some readily accessible design features). The reasons for this may be that involvement of the industry tends to increase the cost of conducting research, and industry people do not want to waste their R&D investment in research that will seem to be suboptimal or of bad quality to a learned outsider or a regulatory agency. Bias is introduced primarily not by poor research design and conduct, but by picking in advance the research questions in a way that the desired results will be obtained, such as by choosing strawmen comparators against which new drugs can definitely be shown to be as good or even better or by selecting outcomes in which the response is easy to see, even though they are not the outcomes that really matter. In contrast to such corporate bias, psychological science seems to be infiltrated mostly by biases that have their origin at academic investigators. As such, they revolve mostly along the axes of confirmation and allegiance biases. Academics may want to show that their theories, expectations, and previous results are correct, regardless of whether this has also any financial repercussions or not.

## Incentives for Replication and for Correcting Wrong Results

There has been little empirical research on ways of correcting wrong results and generally for increasing the credibility of

science. Some suggestions for potential amendments that can be tested have been made in previous articles (Ioannidis, 2005; Young, Ioannidis, & Al-Ubaydli, 2008) and additional suggestions are made also by authors in this issue of *Perspectives*. Nosek et al. (2012) provide the most explicit and extensive list of recommended changes, including promoting paradigm-driven research; use of author, reviewer, editor checklists; challenging the focus on the number of publications and journal impact factor; developing metrics to identify what is worth replicating; crowdsourcing replication efforts; raising the status of journals with peer review standards focused on soundness and not on the perceived significance of research; lowering or removing the standards for publication; and, finally, provision of open data, materials, and workflow. Other authors are struggling with who will perform these much-desired, but seldom performed, independent replications. Frank and Saxe (2012) and Grahe et al. (2012) suggest that students in training could populate the ranks of replicators. Finally, Wagenmakers et al. (2012) repeat the plea for separating exploratory and confirmatory research and demand rigorous a priori registration of the analysis plans for confirmatory research.

All of these possibilities need to be considered carefully. As I have argued earlier, it is essential that we obtain as much rigorous evidence as possible, including experimental studies, on how these practices perform in real life and whether they match their theoretical benefits (Ioannidis, 2012b). Otherwise, we run the risk that we may end up with worse scientific credibility than in the current system. Many of these ideas require a change in the incentives pattern of current science, focusing on truth when we try to reward (or penalize) specific scientific results (Ioannidis & Khoury, 2011). This may not be easy to achieve unless all the major players (scientists, sponsors, journals, industry, regulators, general public) are in line with such a plan, and they expect that science is about getting as close to the truth as possible and not about getting spectacular, but wrong, results.

Some of the proposed steps may even harm the credibility of science unless the pursuit of truth is given the highest priority. For example, crowdsourcing replication and/or using students in training to do the replication work may reinforce the impression that replication is not serious science and that anyone can do it—it is a game of no consequence that is unfit for senior, famous professional scientists. Similarly lowering or removing the standards for publication may result in a flurry of junk with total depreciation of the value of the scientific paper, as any paper without any bare minimum requirements can be published. Propagation of paradigm-driven research may sometimes lead to creation of popular bubbles surrounding bandwagon paradigms. Developing metrics to prioritize replicating certain specific results and not others may give the impression that replication is sparingly useful, whereas the correct message should probably be that replication is useful by default. Checklists for reporting may promote spurious behaviors from authors who may write up spurious methods and design details simply to satisfy the requirements of having done a good study that is reported in full detail; flaws in the design and execution of the study may be buried under such normative responses. Registration practices need careful consideration to decide what exactly needs to be registered: a study, a dataset, or an analysis (Ioannidis, 2012a). For example, just registering a randomized trial without providing details about its analysis plan may offer a false sense of security, whereas data can still be analyzed selectively to get the desirable result. Conversely, registering a detailed analysis a priori does not mean that everything can be anticipated, especially as research involves humans, who are unpredictable: Participants may be lost to follow-up, miss measurements, change treatments, or do weird things that the analysis plan may not have fully expected. Finally, even though open availability of raw data, protocols, and analyses is intuitively the best way to move forward in full openness and transparency, there is a chance that opening up so much data to so many millions of scientists (or people who want to make a living out of science) may promote an exponential growth of free data dredging.

I do not mention all of these caveats because I believe they are very likely to occur to the point that the negatives of these proposed practices will outweigh the positives. I think that adopting some or even all of the system changes proposed in previous articles and in articles in this issue of *Perspectives* will likely do more good than harm. However, at the end of the day, no matter what changes are made, scientific credibility may not improve unless the pursuit of truth remains our main goal in our work as scientists. This is a most noble mission that needs to be continuously reasserted.

## Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

## References

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. (2011). Public availability of published research data in high-impact journals. *PLoS ONE*, *6*, e24357.

Aly, G. (1994). *Cleansing the fatherland: Nazi medicine and racial hygiene*. Baltimore, MD: The Johns Hopkins University Press.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research*, *43*, 666-678. doi:10.3758/s13428-011-0089-5

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives in Psychological Science*, *7*, 543–554.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., & Stodden, V. (2009). Reproducibility research in computational harmonic analysis. *Computing in Science & Engineering*, *11*, 8–18.

Dwan, K., Altman, D. G., Cresswell, L., Blundell, M., Gamble, C. L., & Williamson, P. R. (2011). Comparison of protocols

and registry entries to published reports for randomised controlled trials. *Cochrane Database of Systematic Reviews*, *19*. doi:10.1002/14651858.MR000031.pub2

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, *60*, 411–415.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), 1–11.

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLoS ONE*, *5*, e10271.

Fanelli, D. (2010b). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*, e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120–128.

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171.

Francis, G. (2012a). The psychology of replication and replication in psychology. *Perspectives in Psychological Science*, *7*, 585–594.

Francis, G. (2012b). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives in Psychological Science*, *7*, 600–604.

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives in Psychological Science*, *7*, 639–642.

Galak, J., & Meyvis, T. (2012). You could have just asked: Reply to Francis (2012). *Perspectives in Psychological Science*, *7*, 595–596.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives in Psychological Science*, *7*, 562–571.

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives in Psychological Science*, *7*, 605–607.

Hothorn, T., & Leisch, F. (2011). Case studies in reproducibility. *Briefings in Bioinformatics*, *12*, 288–300.

Hubbard, R., & Armstrong, J. S. (1994). Replication and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*, 233–248.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.

Ioannidis, J. P. A. (2012a). The importance of potential studies that have not existed and registration of observational datasets. *JAMA: Journal of the American Medical Association*, *308*, 575–576.

Ioannidis, J. P. A. (2012b). Scientific communication is down at the moment, please check again later. *Psychological Inquiry*, *23*, 267–270.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., . . . van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*, 149–155.

Ioannidis, J. P. A., & Khoury, M. J. (2011). Improving validation practices in "omics" research. *Science*, *334*, 1230–1232.

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*, 543–549.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Kelly, C. W., Vhase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research: An analysis. *Human Communication Research*, *5*, 338–342.

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science*, *7*, 537–542.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology*, *1*, 115–126.

Merton, R. K. (1973). *The sociology of science, theoretical and empirical investigations*. Chicago, IL: The University of Chicago Press.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, *5*, 85–90.

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

Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives in Psychological Science*, *7*, 643–644

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives in Psychological Science*, *7*, 615–631.

Osherovich, L. (2011). Hedging against academic risk. *Science-Business eXchange*, *4*(15), doi:10.1038/scibx.2011.416

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives in Psychological Science*, *7*, 531–536.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

Proctor, R. N. (2011). *Golden holocaust: Origins of the cigarette catastrophe and the case for abolition*. Berkeley: University of California Press.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature*, 470, 437.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family marital psychotherapy literature. *Clinical Psychology Review*, *9*, 589–603.

Simera, I., Moher, D., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). A catalogue of reporting guidelines for health research. *European Journal of Clinical Investigation*, *40*, 35–53.

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. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives in Psychological Science*, *7*, 632–638.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*, 1418–1422.