Failure to Replicate: Sound the Alarm

By John P.A. Ioannidis, M.D., D.Sc.

Editor's Note: Science has always relied on reproducibility to build confidence in experimental results. Now, the most comprehensive investigation ever done about the rate and predictors of reproducibility in social and cognitive sciences has found that regardless of the analytic method or criteria used, fewer than half of the original findings were successfully replicated. While a failure to reproduce does not necessarily mean the original report was incorrect, the results suggest that more rigorous methods are long overdue.

Psychological science has been a highly prolific discipline. Compared with other scientific fields, it has had one of the highest rates of experimental "success." Analyses have shown that almost all studies in the field (90 to 100 percent) claim statistically significant results with p-values (which indicate the likelihood that the experiment's outcome represents mere statistical "noise") of less than 0.05.[1,2]

This may sound like a cause for celebration: Success seems to be ubiquitous! In fact, it should be a cause for concern. Other analyses have shown that the statistical power of studies in the field is too modest on average[3-5] to account for such a high success rate. In other words, the statistical "noise" inherent in these studies has been so high that it should have caused many more negative results than were reported—even if all the hypotheses targeted in these studies were true.

The lack of replication is even more worrisome. Psychological science has one of the lowest rates of replication studies, in particular exact replications by independent investigators. A recent text-mining survey of the 100 most-cited psychology journals since 1900 found that only 1.07 percent of the published papers were categorized as replications.[6] Some replications may have been missed by the survey authors' automated search, but almost certainly not many. Of the identified replications, only 18 percent were true replications, the remainder being extensions of the work using different methods, settings, populations, or other deviations ("conceptual replication").[6] Furthermore, only 47 percent of the identified replications were done by investigators who were not authors of the original studies.[6]

Such figures reflect psychological science's incentive structure, in which replication experiments have been relatively unwelcome: They have attracted little funding, typically have been harder to get published, and have received little academic recognition.

Several scientists have argued that this is a recipe for disaster.[7-9] Indeed; this author has proposed that in scientific fields where underpowered experiments are the norm and significance-chasing behaviour thrives, one would expect the *majority* of "statistically significant" results to be false-positives.[7]

Conceptual replication can offer complementary insights, but cannot replace direct, exact replication. When there are many unsorted false-positives, conceptual replication with a pressure to find more significant results may simply perpetuate fallacies and lead even more investigators astray.

Replication done by the original authors may also have value, but can lead to a culture of "inbreeding"[10] in which each scientific finding is reproducible only within a restricted environment—the laboratory of a professor and his or her team and mentees. Outsiders who attempt to enter this closed world may "spoil" the results, much like explorers who have entered ancient sealed tombs only to find that beautiful coloured frescoes within are blanched by the contact with fresh air.

A Game Changer?

Theoretical concerns and sporadic evidence have not been able to convince the field to change its dislike for exact replication. However, this pattern may now change, because far more powerful, and hopefully more convincing, evidence has emerged from the Reproducibility Project, led by the Center for Open Science in Charlottesville, Va.

In this four-year project[11], 270 experienced investigators joined forces to conduct exact and adequately powered replications of 100 studies that had been published in three leading psychology journals. The exercise was carried out with exemplary rigour and involved close communication with the original authors to ensure that the replication adhered as faithfully as possible to the original experimental conditions. There are different statistical approaches to define successful replication, but all of these suggested that nearly two-thirds of the original findings were false-positives, with worse performance in social psychology than in cognitive psychology.[11]

For example, one replication study tried to replicate whether participants primed with close spatial distances would report stronger feelings of closeness to their family, siblings, and hometown than participants primed with long distances, as proposed in an earlier paper published in 2008.[12] Despite using identical stimulus materials, dependent variables, and analysis strategies, the replication effort could not replicate the original findings on spatial priming and emotional closeness. Another replication study aimed to replicate that reduced self-regulation resources correlate with increased biases in confirmatory information processing, as previously published.[13] The original paper had shown that the depletion of self-regulation resources influences the search and the processing of standpoint—consistent information in a personnel decision case, even when

confronted with an alternative explanation, i.e. the ego threat, with associated failure cognitions and negative emotions. This could not, however, be documented in the replication effort. More examples and details on the studies replicated can be found in https://osf.io/ezcuj/.

The results of the Reproducibility Project caused a flurry of interest in the scientific community and the general public.

The Reaction

Some of the immediate responses were wrong or counter-productive. On one extreme, commenters suggested that psychology is not a science and should be abandoned or be called an art. On the other extreme, some dismissed the failures to replicate as having been due presumably to unknown differences between the original experimental setups and the replication attempts.

Let's not spend time arguing whether psychology is a science. It is a very important science, and, as the Reproducibility Project reminds us, has been at the forefront of the study of the scientific method and its biases.

Lack of replication and reproducibility has been documented in other scientific disciplines[14-16]. In fact, those that have recently started performing replication experiments have seen very high replication failure rates, even higher than those of the Reproducibility Project in psychology.[17,18] Meanwhile disciplines that have adopted replication in large-scale, e.g., genetic epidemiology, have seen dramatic improvements in the reliability of their results.[19] Many fields of neuroscience and neurobiology are characterized by the conduct of small, underpowered studies[5] and reproducibility

is likely to be low. Thus, what we have documented in the Reproducibility Project may be a pattern that affects many other disciplines.

It is also easy to refute the suggestion that unknown experimental differences are the chief cause of irreproducibility. If that were the case, one would have seen larger as well as smaller effects in the replication studies, compared to the originals. In fact, the replication effect sizes were almost always markedly lower in the replication efforts, rendering them statistically non-significant.

Probably most of the replication failures in psychological science are due to bias in the original results. It is not possible to pinpoint exactly which specific study was biased and how bias exactly happened—replications may also have been biased occasionally. However, the notion that all results are correct despite failures to reproduce them amounts to irresponsible hand-waving. If we want a research finding to make any claim to generalizability, or better yet to be used for practical purposes, other scientists should be able to reproduce it relatively easily. No one would like to fly in a plane that has flown successfully only once, especially if its manufacturers are satisfied that it flew only once and don't mind that it may crash on its second flight. And of what use would a plane be if it flew once and was dismantled, and afterwards no one could rebuild it?

In the Reproducibility Project, one-third of the 147 studies that were identified as possible targets for replication were not picked by any of the 270 replicators, since they were felt to be too difficult, if not impossible, to even try to replicate.[11] It is unclear what the value of research is when no one, other than the original scientists, can ever approach it. Among the studies in the Project for which

replications were made eventually, difficulty in building the replication experiments was a predictor of replication failure.[11]

The Reproducibility Project will hopefully lead to a better appreciation of the need for incorporating exact replication more routinely in the life-cycle of research in psychological science. There is clearly a need for more replication studies, done by independent investigators. There are, however, still many unanswered questions and concerns about how to optimally implement a replication agenda.

Other Considerations

One major concern is the level of resources required. Doing replication well takes a lot of effort. Hastily conceived, suboptimal efforts may even do harm by generating spurious results and confusion. A replication agenda will require substantial funding. While this may be seen as eroding a discovery budget that is already constrained, such a perspective would be misleading. Replication is not some sort of unfortunately imposed policing; it is actually an integral part, perhaps the most integral part, of the scientific discovery process. If the current situation is such that the majority of "discoveries" are false, then replication is the most essential element in any true discovery. Replications also allow us to identify rapidly the avenues of research that warrant further investigation and have the best potential for future yield. In short, more reliance on replication can help save us from fund-wasting dead-ends and false-positives.

Should everything be subjected to replication? Some other scientific fields have accepted this as a norm. In genetic epidemiology, for example, it is impossible to publish anything in a high-profile journal without independent replication. However, in the field of psychology there may be

insurmountable barriers to the adoption of this principle. These include practical difficulties (as discussed above, e.g. for very complex experiments). Also, the community may not be ready for such a sweeping paradigm shift. It may be necessary to target replication efforts in a more limited, strategic way.

As a first step, research could be categorized as "replicated" or "unchallenged."[20] Unchallenged research would have to be treated with extra caution—as more likely false than true, perhaps with substantial variability across sub-disciplines. Samples of studies of different types, and stemming from different sub-disciplines, would be subjected to replication periodically to examine what is the current replication performance of the sub-discipline. Therefore one would know that working in field X with study design Y carries a Z percent risk of non-replication. Such figures would change over time, particularly if the field were to adopt more safeguards to improve its overall research practices. These safeguards could include registration of research protocols prior to experiment, data sharing, team science approaches or other practices that improve transparency, efficiency, and reliability.[21-23]

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.

Finally, studies that aim to inform practical applications or otherwise affect humans, such as treatments for psychological problems, should have thorough replication as a *sine qua non*, before being adopted in everyday practice.

At a first stage, such a replication science agenda is also likely to require a very small amount of funds, perhaps 3 to 5 percent of the current research budget—a bargain if it reduces the 50 to 90 percent of the research budget that currently seems to be wasted on irreproducible research. That said, the devil can be in the details, such as who will fund replications, when should they take place, and how should they be conducted. Editors and reviewers also need to become friendly to good replications[24,25]; publishing replications will only greatly encourage replication.

To get where we need to go, all action plans will need to have strong grass-roots endorsement by the scientific community. The Reproducibility Project, and the favorable responses to it, show that many scientists care deeply about making research more reproducible. There is no reason to doubt that the general public would also want the same.

Bio

John P.A. Ioannidis, M.D., holds the C.F. Rehnborg Chair in disease prevention at Stanford University; is professor of medicine, and of health research and policy; and director of the Stanford Prevention Research Center at the School of Medicine; professor of statistics (by courtesy) at the School of Humanities and Sciences; one of the two directors of the Meta-Research Innovation Center; and director of the Ph.D. program in epidemiology and clinical research. Ioannidis, who grew up in Athens, Greece, received a doctorate in biopathology from the University of Athens and trained at Harvard and Tufts (internal medicine and infectious diseases), then held positions at NIH,

Johns Hopkins, and Tufts. He has served as president of the Society for Research Synthesis Methodology.

References

1. M. Bakker, A. van Dijk, and J.M. Wicherts, "The Rules of Game Called Psychological Science," Perspectives on Psychological Science 7 (2012): 543-54.
2. D. Fanelli, ""Positive" Results Increase Down the Hierarchy of the Sciences," PLoS ONE 5 (2010): e10068.
3. S.E. Maxwell, "The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies." Psychological Methods, 9 (2004): 147-163.
4. J.P. Ioannidis, M. Munafò, P. Fusar-Poli, B.A. Nosek, S. David, "Publication and Other Reporting Biases in Cognitive Sciences: Detection, Prevalence and Prevention" Trends in Cognitive Sciences 18 (2014): 235-41.
5. K.S. Button, J.P. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S. Robinson, and M.R. Munafò, "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," Nature Reviews Neuroscience 14 (2013): 365-76.
6. M. Makel, J. Plucker, and B. Hegarty, "Replications in Psychology Research: How Often Do they really Occur?" Perspectives on Psychological Science 6 (2012): 537-42.
7. J.P. Ioannidis, Why Most Published Research Findings are False. PLoS Medicine 2 (2005): e124.
8. K. Fiedler, "Voodoo Correlations are Everywhere. Not Only in Neuroscience," Perspectives on Psychological Science 6 (2011): 163-171.
9. B. A. Nosek, and Y. Bar-Anan, "Scientific Utopia: I. Opening Scientific Communication," Psychological Inquiry 23 (2012): 217-43.

10. J.P. Ioannidis, “Scientific Inbreeding and Same-Team Replication,” Journal of Psychosomatic Research 73 (2012): 408-10.
11. Open Science Collaboration, “Estimating the Reproducibility of Psychological Science,” Science 349 (2015): aac4716.
12. L.E. Williams, J.A. Bargh, J. A, “Keeping One’s Distance: The Influence of Spatial Distance Cues on Affect and Evaluation,” Psychological Science 19 (2008): 302-8.
13. P. Fischer, T. Greitemeyer, D. Frey, “Self-regulation and Selective Exposure: The Impact of Depleted Self-regulation Resources on Confirmatory Information Processing,” Journal of Personality and Social Psychology 94(2008): 382-395.
14. H. Evanschitzky, C. Baumgarth, R. Hubbard, and J.S. Armstrong, “Replication Research’s Disturbing Trend,” Journal of Business Research 60 (2007): 411-5.
15. R. Hubbard, and J.S. Armstrong, “Replication and Extensions in Marketing: Rarely Published but Quite Contrary,” International Journal of Research in Marketing 11 (1994): 233-48.
16. C.W. Kelly, L.J. Vhase, and R.K. Tucker, “Replication in Experimental Communication Research: an Analysis,” Human Communication Research 5 (1999): 338-42.
17. C.G. Begley, and L.M. Ellis, “Drug Development: Raise Standards for Preclinical Cancer Research,” Nature 483 (2012): 531-3.
18. F. Prinz, T. Schlange, and K. Asadullah, “Believe it or not: How Much can we Rely on Published Data on Potential Drug Targets?” Nature Reviews Drug Discovery 10 (2011): 712-713.
19. J.P. Ioannidis, R. Tarone, and J.K. McLaughlin, “The False Positive to False Negative Ratio in Epidemiologic Studies,” Epidemiology 22 (2011): 450-6.
20. J.P. Ioannidis, “Why Science is not Necessarily Self-correcting,” Perspectives in

Psychological Sciences 7 (2012): 645-54.

21. J.P. Ioannidis, "How to Make More Published Research True," PLoS Medicine 11 (2005): e1001747.
22. D.L. Donoho, A. Maleki, I.U. Rahman, M. Shahram, and V. Stodden, V, "Reproducibility Research in Computational Harmonic Analysis," Computing, Science, & Engineering 11 (2009): 8-18.
23. J.M. Wicherts, D. Borsboom, J. Kats, and D. Molenaar, "The Poor Availability of Psychological Research Data for Reanalysis," American Psychologist 61 (2006): 726–728.
24. J.W. Neuliep, and R. Crandall, "Editorial Bias Against Replication Research," Journal of Social Behavior and Personality 5 (1990): 85–90.
25. J.W. Neuliep, and R. Crandall, "Reviewer Bias Against Replication Research," Journal of Social Behavior and Personality 8 (1993): 21–29.