books & ideas

Non-replicable

The methodological crisis in experimental science

by Aurélien Allard

Over the last twenty years, some of the social and bio-medical sciences have had to face the impossibility of replicating some of their most famous experiments. Although this situation raises serious concerns, it is an opportunity to fundamentally renew the scientific methodology in these disciplines.

The social and bio-medical sciences are currently facing one of the most important crises in their history. Since the beginning of the 2000s, and particularly since 2011, it has become increasingly clear that numerous experiments in the fields of psychology, economics or animal studies, are quite simply not replicable, and do not represent true scientific knowledge. The methods employed today in these disciplines face mounting criticism for their lack of rigour. An increasing number of researchers are calling for a radical renewal of the scientific practices in these fields.

Numerous international scientific collaborations seeking to replicate past experiments have been enacted over recent years. Replicating an experiment or a survey consists in attempting to reproduce it identically in order to discover to what extent similar results can be obtained. Between 2012 and 2015, a large-scale international collaboration attempted to evaluate the replicability of social and cognitive psychology experiments and surveys. Out of the 100 experiments the *Reproducibility Project: Psychology* attempted to reproduce, only 39 showed results that conformed to those obtained in the original experiment. These results were widely discussed, both within the research community as well as in the general Anglo-American press. In the field of experimental economics, a similar project, conducted on a smaller scale, only succeeded in replicating the results of 11 of the 18 experiments attempted. A comparable effort is ongoing in the field of cancer biology. Between 2011 and 2012, researchers working for private companies (*Bayer* and *Amgen*) warned the scientific community that the vast majority of experiments published in the bio-medical disciplines

were not replicable. They reported that they had only managed to replicate a tiny minority (between 11 and 25%)[1] of the experiments.

While these results reveal certain crucial problems in the social and bio-medical sciences, they also represent an opportunity for a radical improvement in the research practices in these fields. The existence of international collaborations reveals both the scale of the problem of replicability and the desire of the reform movements to improve the existing practices. There has been an unprecedented increase in methodological reflection and projects to improve practices over the last years. Certain commentators thus see the 2010s as a period that provoked a true renaissance of psychology as a scientific discipline.

Reproducing and replicating

We could believe that the issue of replicability primarily concerns the experimental disciplines. The general question of the possibility of reproducing an observation is nonetheless fundamental for any discipline with scientific pretensions, regardless of whether it is qualitative or quantitative, experimental or not. To understand the issues linked to the question of replicability, we must make a distinction between the analytical reproduction of an observation, and its replication.[2]

Analytically reproducing an observation consists in seeking to reproduce the results using the materials originally used by the experimenter. For example, in the case of qualitative research based on interviews, for example in the fields of anthropology or sociology, reproducing the results could consist in consulting recordings made by the anthropologist. This is a means of checking whether the people who participated in the survey actually said what the anthropologist stated they did in his or her research. In the case of quantitative research, the same dataset used by the original researcher can be used to carry out the same statistical analyses, and to check that the results correspond. An analytical reproduction of the results is crucial, both to forestall the problem of scientific fraud, and to identify the honest mistakes that can lead scientists to report results contradictory to their findings.

Although the possibility of reproducing the results of research is an important issue, the problems raised in this article mainly concern the question of replicability. The latter

[1] Regarding these different replication projects, see respectively Open Science Collaboration, 'Estimating the Reproducibility of Psychological Science', *Science*, 349, 2015; Colin Camerer et al., 'Evaluating Replicability of Laboratory Experiments in Economics', *Science*, 2016 ; C. Glenn Begley and Lee M. Ellis, 'Drug Development: Raise Standards for Preclinical Cancer Research', *Nature*, 483, 2012; Florian Prinz, Thomas Schlange, & Khusru Asadullah, 'Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets?', *Nature Reviews Drug Discovery*, vol. 10, 2011, 712.

[2] In English, the distinction between replicability and reproducibility is somewhat vague. Various authors employ these terms in sometimes contradictory manners. The distinction adopted here is nonetheless a common one. See Roger D. Peng, Francesca Dominici, Scott L. Zeger, "Reproducible Epidemiologic Research", *American Journal of Epidemiology*, vol.163, n° 9, 2006, pp. 783-789.

concept can be applied to any situation where a researcher wants to generalise beyond the cases s/he has observed. Replicating an observation hence consists in using the same methods as those employed by the original author, and attempting to see to what extent the fresh observations are compatible with those originally reported. In anthropology, we can recall the famous controversy around Derek Freeman's failure to replicate Margaret Mead's fieldwork in Samoa. While Mead had described Samoa as a society of free sexuality, Freeman had emphasised its puritanical character. Mead had primarily based her fieldwork on interviews with two informants, and had sought to generalise her data to include the sexuality of the whole of Samoa's population. Again in a non-experimental context, to replicate quantitative observations, a new survey, using the same methods, can be conducted in order to measure the extent to which the results can be generalised. For example, Fabien Jobard and Sophie Nevanen observed that ethnic group had no impact on criminal sentencing in a regional court of the Paris region. They then generalised their results, interpreting them as a sign of lack of discrimination throughout France.[3] Replicating these results could take the form of another survey in several other French courts, to see whether it is possible to obtain the same results.

A crisis of confidence

While the issue of replicability concerns all the social and bio-medical sciences, social and cognitive psychology has been central to the crisis and the reform movements.

The 2000s were a halcyon period for social psychology in Great Britain and North America, marked by the publication of numerous best-sellers that awakened the interest of a wide audience.[4] From 2011 onwards, however, a series of scandals revealed certain deep-seated problems that affected the discipline.

In 2011, Daryl Bem, a professor at the prestigious Cornell University in the United-States, published a series of experiments in the *Journal of Personality and Social Psychology*, one of the best-reputed journals in the field.[5] This series of experiments sought to demonstrate the existence of extra-sensory powers: D. Bem attempted to show that humans were capable of foretelling the future.[6] One particular experiment sought to prove that people who can revise *after the exam is over* obtain better marks in this exam than those who do not have this option. In 10 different experiments, D. Bem systematically reported results that conformed to his predictions.

---

[3] Fabien Jobard and Sophie Névanen, 'La couleur du jugement', *Revue française de sociologie*, 2007/2, vol. 48, pp. 243-272.

[4] Among the numerous best-sellers of the 2000s, we can mention the work by Carol Dweck, *Mindset (*Ballantine Books, 2007), or *Stumbling on Happiness* (Random House Books, 2006) by Daniel Gilbert.

[5] Daryl J. Bem, 'Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect', *Journal of Personality and Social Psychology, 100*(3), 2011, pp. 407-425.

[6] Although this may seem like a hoax, Bem was perfectly serious and truly believed in extra-sensory powers.

It goes without saying that such results contradict all the known physical laws and almost no one believed them at the time. Nonetheless, these experiments raised a key issue for social and cognitive psychology, because the methods Daryl Bem employed were exactly the same as those employed in research that had been considered perfectly respectable until this point. The existence of this type of results was highly problematic as it proved that it was possible to 'prove' anything, including absurd conclusions, using experimental methods traditionally employed in the field of psychology. In the following years, several articles naturally failed to replicate D. Bem's results.

While the publication of extra-sensory experiments came as a great shock to the scientific community, a series of similar scandals led its members to question themselves at a much deeper level. Again in 2011, a reputed professor from the University of Amsterdam, Diederik Stapel, was dismissed for having invented his own experimental results. In the 2010s, several classical social psychology experiments failed the test of replicability. All these problems gave rise to a wide reform movement, which in 2015 culminated in the publication of the *Reproducibility Project: Psychology* (*RP:P*), mentioned in the introduction.

The *RP:P* hence suggests that the majority of psychology experiments greatly exaggerate the importance of the results studied. The effects reported in the original articles were, indeed, on average twice as important as those obtained in the replications. Of course, this does not mean that all research in the field of psychology is skewed. In 2014, a project to reproduce 13 social and cognitive psychology experiments, the *Many Labs Project*, produced contrasting results in terms of replicability. The *RP:P* experiments had been chosen with an aim to obtain a representative sample of the discipline. Adopting a different approach, the *Many Labs Project* carefully chose classical psychology experiments in order to study the extent to which it was possible and easy to reproduce these results, in other laboratories, working with diverse populations. The attempts at replication were conducted in 9 different countries and involved 10 classical experiments in the discipline and three more recent ones. While the research team managed to reproduce the 10 classical experiments, none of the three recent experiments passed the test of replicability.[7] A majority of psychology experiments might be unexploitable, but a nucleus of reliable results certainly exists within the discipline.

Other social sciences or bio-medical disciplines face similar difficulties. The case of bio-medical research is complex as fairly rigorous fields coexist with other areas where there is a clear lack of precision. In short, there is a massive contrast between the research conducted on human beings, which is generally of quite high quality, and that involving animals, which is often of very dubious quality. Medical research conducted on humans is regulated by relatively strict rules that force it to respect fairly strong scientific demands, particularly when

[7] Richard A. Klein et al., 'Investigating Variation in Replicability: A "Many Labs" Replication Project', *Social Psychology, 45* (3), 2014, p. 142-152.

it comes to tests conducted prior to the commercialisation of a new drug.[8] The preparation for commercialisation involves several phases of tests that seek to eliminate medicines that may be dangerous for humans, or that have not proven their efficiency. Inversely, research carried out on animals is subject to few regulations and the research practices in this area are often mediocre.[9]

This situation leads to a large amount of financial wastage, given that drugs tested on humans are based on studies conducted on animals. If a molecule seems to have shown promise in treating a cardiac problem in mice, a medicine based on this molecule will probably be developed and tested on humans. However, as the initial research on animal populations is not conducted very rigorously, most of the time the studies involving people fail to show that the medicine in question can provide the slightest benefit. Thus a mere 5% of the medicines tested on humans turn out to prove beneficial and are commercialised.[10]

Why is research so unreliable?

We can identify four key factors that make the results published in social science and bio-medical journals unreliable: the fact that scientific culture only focuses on innovative and 'positive' results, excessive flexibility in the statistical analysis of the results, the lack of replication of earlier works carried out by laboratories other than those that had published the original studies, and the weakness of the sample sizes.

Scientific culture in the social and bio-medical sciences strongly valorises innovative and positive results. Scientific journals would prefer to publish articles dealing with a new phenomenon, particularly if they show that this phenomenon may have practical and decisive applications. The main consequence of this approach is that researchers are encouraged not to publish 'negative' results, or results that do not show a significant difference between various treatments. For example, in the medical field, an article dealing with a promising cancer treatment is more likely to be published than an article describing how a new treatment fails to cure it. As a result, only some experiments are published, which leads to in what is known as 'publication bias'.

This kind of bias can result in the exclusive publication of articles that show a difference, even when none exists. Imagine 20 medical laboratories seeking to determine whether a specific molecule can cure a cancer. Inevitably, purely by chance, certain

---

[8] This is an over simplification: the problems of replicability are also highly pregnant in studies dealing with human health, although the situation is far better than in the case of animal studies.
[9] David W. Howells, Emily S. Sena, & Malcolm R. Macleod, 'Bringing Rigour to Translational Medicine', *Nature Reviews Neurology*, 10, 2014, pp. 37–43.
[10] John Arrowsmith, 'A Decade of Change', *Nature Reviews Drug Discovery*, 11, 2012, pp. 17–18.

laboratories will find a positive correlation between this treatment and a cure for the cancer (for example, because the people treated had a cancer that was less aggressive than those in the control group). Others may discover a negative correlation that leads them to believe the treatment aggravates the situation. Ultimately, only the laboratory that showed that the treatment has a significant curative effect on cancer will be able to publish its article, because this is the only one that will be sufficiently interesting. The 19 other laboratories, on the contrary, will probably leave their study to gather dust at the bottom of a drawer, because they know that if the study does not promise to cure cancer, it is not publishable.

The consequence of such practices is that research published in scientific journals fundamentally exaggerates the efficiency of medical treatments, pedagogic or educational interventions, and in the field of psychology, the influence of subtle manipulations on human behaviour. Scientific literature is hence dotted with false, or at least, highly exaggerated results.

This publication bias is all the more problematic as it tends to encourage researchers to 'fix' their results. To maximise the likelihood of being published, they may resort to questionable and less rigorous research practices, which may enhance the probability of obtaining a positive result. A first questionable practice consists of recruiting participants progressively, in several stages, with the intent to stop the experiment only when the desired results are obtained. A second common practice is to test several parameters simultaneously, in an attempt to compare the multiple possible results before focusing on the only experiment that 'worked'. A third type of questionable practice is to conduct analysis only on a sub-group of participants. For example, in bio-medical research, if some researchers obtain results indicating that a new medicine is not very efficient, they may attempt to subdivide their analyses to see whether they can find a positive effect on women, men, young people or old people, etc. By dividing the population up into multiple sub-groups, there is a strong likelihood that, purely by chance, they may find that the medicine is beneficial to a particular group, even if, in reality, it is beneficial to no one.

This lack of reliability in the initial research would not be major problem in itself if scientific research corrected itself over time. In the scientific world, this could take the form of attempts to replicate earlier experiments. This type of practice is very common in physics, for example. As we saw, replicating an earlier experiment has become a fairly common practice in the field of psychology since the beginning of the replicability crisis in 2011. But this practice had been almost unheard of until that time: it was extremely difficult to have a replication published, as those articles were not considered sufficiently innovative. As a result, completely erroneous studies continued to be cited for decades, and were never corrected.[11]

[11] Among the examples of non-replicable experiments that have had an enormous impact, we can, for example, mention the experiment conducted by John Bargh and his colleagues in 1996, that attempted to show that presenting words associated with old age lead young subjects to behave like old people, for example, walking slowly. The failure of the attempt at replication by Doyen and his colleagues in 2012 was one of the factors at

Last but not least, the lack of reliability is also caused by the weakness of the sample sizes. The experiments are indeed carried out on too few subjects. Now, the larger the sample size, the higher the precision in the evaluation of the scientific results. In the case of animal research, people often study the efficiency of treatments for cardiac problems on fewer than 10 animals, including the control group. It is difficult to imagine how research on such a low number of subjects could produce results that can be generalized. While in the field of psychology research on such a small sample is rare, the comparison of groups of twenty or thirty people was the norm for years. It is true that groups of this size can suffice to show the existence of very strong psychological effects. The *Stroop* effect, as this condition is called, shows the difficulty subjects have in pronouncing a word representing a colour (red, for example) if it is written in an ink of a different colour (in green, for example). This effect is widespread, highly robust and easy to demonstrate with a very low number of participants. But any rigorous demonstration of more subtle effects, — which probably constitute most of the effects studied in psychology — is quite simply impossible with such low sample sizes.

What can be done? The current methodological reform

Given the magnitude of the problem of replicability, the research community in the field of psychology has recently begun to enact deep methodological reforms at an unprecedented scale in the social sciences. The first impact, mentioned above, is the multiplication of replication projects. While this reform is crucial, it is not the only one.

A second reform trend seeks to increase the precision of experimental measurements largely by increasing sample sizes. Although there has been a great deal of support for this project, the progress so far remains limited. One of the main obstacles to conducting large-scale experiments, of course, lies in the cost of such studies; few laboratories are, in fact, able to have 1000 different participants undergo the same psychological experiment. To overcome this obstacle different methods of recruitment, on Internet, have emerged over the last few years. Thus a large number of recently conducted surveys have been conducted via a service offered by Amazon, *Amazon Mechanical Turk*, which makes it possible to recruit a large number of participants for a fee. While this system indeed helps increase the number of participants, the representativeness of the recruited sample should be questioned: the participants are mainly Indian or American, and in the case of the United States, they are generally slightly younger, better educated and politically further to the left than the overall

---

the origin of the replicability crisis. See S. Doyen, O. Klein, C-L. Pichon, A. Cleeremans, 'Behavioral Priming : It's All in the Mind, but Whose Mind?', *PLoS ONE*, 2012, 7(1): e29081.

American population.[12] However, all experiments cannot be carried out at a distance. A major proposal is to develop collaborations between laboratories in order to share resources. Developments of this type have helped to improve the reliability of genetic research: while the work was mediocre in quality until the beginning of the 2000s, the development of international collaborations has brought about a drastic improvement in the quality of the research in this field. Two international collaborative projects, in the fields of psychology and neuroscience, were thus implemented in 2017.[13]

A third key aspect of the ongoing methodological reform seeks to remedy the problems raised by the lack of rigour applied to data analysis. As we saw, researchers generally have a lot of room for manoeuvre in data analysis, which makes it possible to arrive at a favourable interpretation of the results obtained during an experiment, at a later stage. As a remedy against this, the idea of preregistration is gaining traction. This method forces the researcher to announce the hypothesis he wants to test, and the manner in which the data will be analysed, before beginning the experiment. The preparatory publication is then uploaded to an Internet site that freezes the document, preventing any further modifications. After the experiment has been conducted the researcher has to allow the results to speak for themselves, and s/he cannot carry out a questionable analysis that *in fine* confirms her or his theory. Preregistration is a common practice in medical studies applicable to humans; applying it to the social sciences and animal studies would be real progress.[14]

Following the same reasoning, an increasing number of scientific journals allow researchers to submit their articles before conducting the experiment. This new publication method, know as '*registered reports*' allows peer-reviewers to evaluate the experiment solely on the basis of the method, independently of the nature of the results obtained, whether they be positive or not. A large number of social and cognitive psychology journals have also begun to offer this mode of publication.

These reform movements only represent some of the proposals put forward to improve the reliability of scientific publications. New organizations, like the *Center for Open Science* or the *Society for the Improvement of Psychological Science* were created to accelerate the movement. Such calls for reform have, of course, provoked a certain reticence, or even opposition within

---

[12] The majority of psychology articles are published based on studies of American participants, whether they be students, or members of *Amazon Mechanical Turk*. Beyond the representativeness of the participants from *Amazon Mechanical Turk* in comparison to the general American public, it is of course hard to know the extent to which it is possible to develop universal psychological theories on the basis of the inhabitants of a single country. Over the last years, there have been numerous calls to develop an intercultural psychology. See J. Henrich, S. Heine, and A. Norenzayan, "The Weirdest People in the World ?", *Behavioral and Brain Sciences*, 2010, 33 (2-3), pp. 61-83.

[13] Dalmet Singh Chawla, "A New 'Accelerator' Aims to Bring Big Science to Psychology", *Science*, 2017.

[14] We should nonetheless note that the practice of preregistration is not always ideal. On the one hand, numerous predictions are often quite vague, and this leaves room for manoeuvre in the analysis of the results. On the other hand, researchers sometimes lie about their predictions in order to increase their chances of publishing the article in question. In the latter case, preregistration nonetheless makes it possible to prove that the author of the article surreptitiously carried out analyses that contradict his/her own predictions.

the scientific community. Some voices have been raised in defence of the traditional methods of practicing psychology and the quantitative social sciences. For example, these partisans of the status quo defend the idea that a flexible interpretation in the analysis of results encourages researchers' creativity, and this may be strangled by the excessive rigour demanded by the reformers. Other researchers have also underscored the fact that the money and time dedicated to carrying out replications of earlier experiments would be better used if it were spent on research in new areas. While such criticisms are not absurd, they are nonetheless difficult to accept in a context where social science research often finds it difficult to produce credible results.

Further Reading

General works on the topic (psychology and bio-medical sciences)

- Chris Chambers, *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*, Princeton University Press, 2017.
- Richard Harris, *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*, Basic Books, 2017.

Articles and statistical resources

- Daniël Lakens's blog is an influential and helpful resource on issue of replicability (http://daniellakens.blogspot.ca/).
- Leif D. Nelson, Joseph Simmons et Uri Simonsohn, 'Psychology's Renaissance', *Annual Review of Psychology*, 69, 2018.
- Joseph P. Simmons, Leif D. Nelson and Uri Simonsohn, 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', *Psychological Science*, 22, 2011, pp. 1359–66.

Published in laviedesidees.fr, 20 March 2018. Translated from the French by Renuka George with the support of the Florence Gould Foundation.

Published in *Books & Ideas*, 28 May 2018.