Opinion

EMBO reports

Reproducibility crisis in science or unrealistic expectations?

Thiago FA França[1] & José Maria Monserrat[2]

Science appears to be in a crisis caused by the failure to replicate published results, which is undermining confidence in the scientific literature. This reproducibility crisis is not only evident in large-scale replication efforts to evaluate studies from various laboratories [1], but also within laboratories themselves [2]. The problem has been extensively discussed among the scientific community, as many scientists have had troubles with replication themselves [3]. A recent survey of 1,576 researchers found that around 90% agreed that the reproducibility crisis is real [3].

There are many factors that influence the reproducibility of published results, including statistical methods, publication bias, lack of transparency, insufficient detail in the description of methods and variability of conditions and protocols between laboratories [1,2,4]. Some of these, such as lack of transparency and insufficient detail, directly impair scientists' ability to properly evaluate a study and to replicate the experiments. Others, such as misuse of statistics and publication bias, lead to systematic biases that undermine the reliability of the entire literature. But there is another, even more insidious factor that has received considerably less attention: random variation.

A few years ago, Halsey *et al* [5] published a commentary on the sample-to-sample variability of the *P*-value. They showed that with low statistical power—which is ubiquitous in the literature—the *P*-value can vary widely between samples of the same populations. As part of their argument, the authors performed a simulation in which samples of 10 individuals were taken from two populations whose means differed by half a standard deviation—that is, a standardized mean difference of 0.5. The use of a low sample size led to high variability between samples: In one of their comparisons, the difference between the means of the samples was 1.46 standard deviations, while in another, the difference was −0.08. And yet, both samples came from the same populations and were obtained through a simulation of a perfect data collection–all conditions were equal between the "experiments". It is interesting to imagine what would happen if two groups of researchers performed the same experiments, got similar results, and published these—it would likely start a discussion over the cause of such an enormous difference. What have they done different from each other? Why could one not replicate the findings of the other? How can we reconcile such disparate pieces of evidence?

This example demonstrates that even if we conducted two studies under identical conditions, it would still not guarantee similar results, at least not with low sample sizes—and consequently low statistical power—that are commonly found in the literature. It is therefore likely that at least part of the failure to replicate published results is caused by random variability intrinsic to any sampling procedure. Although a lot has been said about the over reliance on *P*-values—which has been portrayed as one of the great villains in the reproducibility crisis [5,6]—and in favor of more informative estimation statistics [5,6], abandoning the *P*-value alone will not solve the problem of variability between studies, as effect sizes and confidence intervals fluctuate as much as *P*-values.

It remains of course important to analyze other factors that may cause systematic differences between repetitions of the same experiment, but it is also necessary to adjust our expectations about the reproducibility of individual studies. Ethical and practical constraints often impose the necessity to work with relatively small samples sizes; in such cases, failure to replicate should not necessarily cast doubt over the validity of the experiment. It is unreasonable to expect that results from a single study can give great predictive power over the results of future experiments. If the fight against the reproducibility crisis aims at improving reliability of individual studies, then we must understand the limits of such reliability in the first place.

Low statistical power is not necessarily a problem *per se;* it only becomes a problem when studies are not replicated. Once we eliminate all causes of systematic bias in the literature, such as publication bias, differences in protocols, and misuse of statistics, replication of published studies will still not yield identical results, but will yield *comparable* results. Such results can be combined using statistical methods, such as meta-analysis [7], to increase the statistical power and provide a more precise and reliable estimate of the effect being studied. As should be expected from the very nature of statistics, answers do not come from individual studies, only from groups of studies. It is not enough to eliminate biases: Replication will always be paramount.

The intrinsic variability between studies has also repercussions for the ongoing replication efforts. In general, replication tries to apply the same protocol and to minimize any differences between experiments; some replication efforts go to great lengths in order to eliminate all sources of variability [2]. A certain degree of similarity is important to guarantee that results are comparable, but too much similarity can become a problem too. Scientists are increasingly concerned that

[1] Programa de Pós-graduação em Ciências Fisiológicas, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil. E-mail: tfafranca@furg.br
[2] Instituto de Ciências Biológicas, Universidade Federal do Rio Grande (FURG), Rio Grande, RS, Brazil

DOI 10.15252/embr.201846008 | EMBO Reports (2018) 19: e46008 | Published online 25 April 2018

too strictly controlled conditions and invariant animal models may impair the ability to generalize from research findings [8]. The results may well hold true only for homogeneous animal strains and the very strict experimental conditions—in other words, strictly controlled conditions generate true findings, but not robust ones. Robust findings that reflect biological heterogeneity require not only replication but also variability.

In the real world, just as in sampling simulations, there is always variation between samples owing to the fact that the individuals being sampled—patients in a clinical trial, mice in a biomedical study, or cells in an *in vitro* study—show some variability in whatever parameter is being analyzed. It is interesting to consider whether the difficulties of reproducing results differ between studies using different kinds of population. For example, we would expect more heterogeneity between patients in a clinical study than between mice with a homogenous genetic background housed under similar conditions. In fact, even if we used more heterogeneous animal models–for example, rodents with different genetic backgrounds and of both sexes—and design multi-laboratory experiments to allow for some variability in experimental conditions, a multi-site clinical trial with subjects from different neighborhoods and lifestyles would still be expected to show more variability. In practice, however, it is very hard to evaluate differences in reproducibility caused by natural heterogeneity, not only because *in vitro* studies, animal studies, and clinical trials tend to have different sample sizes on average and different variability of the specific population of study, but also because estimates of reproducibility of research can vary considerably [9].

It is important to note that this variability in estimates of reproducibility makes it hard to evaluate the crisis itself—the truth is that we do not really know how severe it is. When we combine this uncertainty with unrealistic expectations about the reproducibility of individual studies, it may inflate our perception of the crisis. As has been noted before [10], there are cases where findings were discredited after another laboratory failed to replicate the results. But if the first study was not reliable, what makes the second one more trustworthy? A better approach would be quantitatively combining the two instead of qualitatively putting them against each other. As long as we hold on to unrealistic expectations, even replication efforts my not help us to get closer to the truth.

References

1. Hunter P (2017) *EMBO Rep* 18: 1493 – 1496
2. Lithgow GJ, Driscoll M, Phillips P (2017) *Nature* 548: 387 – 388
3. Baker M, Penny D (2016) *Nature* 533: 452 – 454
4. Wasserstein RL, Lazar LA (2016) *Am Stat* 70: 129 – 133
5. Halsey LG, Everett DC, Vowler S *et al* (2015) *Nat Methods* 12: 179 – 185
6. Claridge-Chang A, Assam PN (2016) *Nat Methods* 13: 108 – 109
7. Borenstein M, Hedges LV, Higgins JPT *et al* (2009) *Introduction to meta-analysis*. Hoboken, NJ: John Wiley & Sons
8. Voelkl B, Vogt L, Sena ES *et al* (2018) *PLoS Biol* 16: e2003693
9. Freedman L, Cockburn IM, Simcoe TS (2015) *PLoS Biol* 13: e1002165
10. Baker M, Dolgin E (2017) *Nature* 541: 269 – 270