INFECTION AND IMMUNITY, Dec. 2010, p. 4972–4975 Vol. 78, No. 12
0019-9567/10/$12.00 doi:10.1128/IAI.00908-10

EDITORIAL
Reproducible Science[▽]

The reproducibility of an experimental result is a fundamental assumption in science. Yet, results that are merely confirmatory of previous findings are given low priority and can be difficult to publish. Furthermore, the complex and chaotic nature of biological systems imposes limitations on the replicability of scientific experiments. This essay explores the importance and limits of reproducibility in scientific manuscripts.

> "Non-reproducible single occurrences are of no significance to science."
> —Karl Popper (18)

There may be no more important issue for authors and reviewers than the question of reproducibility, a bedrock principle in the conduct and validation of experimental science. Consequently, readers, reviewers, and editors of *Infection and Immunity* can rightfully expect to see information regarding the reproducibility of experiments in the pages of this journal. Articles may describe findings with a statement that an experiment was repeated a specific number of times, with similar results. Alternatively, depending upon the nature of the experiment, the results from multiple experimental replicates might be presented individually or in combined fashion, along with an indication of experiment-to-experiment variability. For most types of experiment, there is an unstated requirement that the work be reproducible, at least once, in an independent experiment, with a strong preference for reproducibility in at least three experiments. The assumption that experimental findings are reproducible is a key criterion for acceptance of a manuscript, and the Instructions to Authors insist that "the Materials and Methods section should include sufficient technical information to allow the experiments to be repeated."

In prior essays, we have explored the adjectives descriptive (6), mechanistic (7), and important (8) as they apply to biology, and experimental science, in particular. In this essay, we explore the problem of reproducibility in science, with emphasis on the type of science is that routinely reported in *Infection and Immunity*. In exploring the topic of reproducibility, it is useful to first consider terminology. "Reproducibility" is defined by the *Oxford English Dictionary* as "the extent to which consistent results are obtained when produced repeatedly." Although it is taken for granted that scientific experiments should be reproducible, it is worth remembering that irreproducible one-time events can still be a tremendously important source of scientific information. This is particularly true for observational sciences in which inferences are made from events and processes not under an observer's control. For example, the collision of comet Shoemaker-Levy with Jupiter in July 1994 provided a bonanza of information on Jovian atmospheric dynamics and *prima facie* evidence for the threat of meteorite and comet impacts. Consequently, the criterion of reproducibility is not an essential requirement for the value of scientific information, at least in some fields. Scientists studying the evolution of life on earth must contend with their inability to repeat that magnificent experiment. Gould famously observed that if one were to "rewind the tape of life," the results would undoubtedly be different, with the likely outcome that nothing resembling ourselves would exist (12). (Note for younger scientists: it used to be fashionable to record sounds and images on metal oxide-coated tape and play them back on devices called "tape players.") This is supported by the importance of stochastic and contingent events in experimental evolutionary systems (4).

Given the requirement for reproducibility in experimental science, we face two apparent contradictions. First, published science is expected to be reproducible, yet most scientists are not interested in replicating published experiments or reading about them. Many reputable journals, including *Infection and Immunity*, are unlikely to accept manuscripts that precisely replicate published findings, despite the explicit requirement that experimental protocols must be reported in sufficient detail to allow repetition. This leads to a second paradox that published science is assumed to be reproducible, yet only rarely is the reproducibility of such work tested or known. In fact, the emphasis on reproducing experimental results becomes important only when work becomes controversial or called into doubt. Replication can even be hazardous. The German scientist Georg Wilhelm Reichmann was fatally electrocuted during an attempt to reproduce Ben Franklin's famous experiment with lightning (1). The assumption that science must be reproducible is implicit yet seldom tested, and in many systems the true reproducibility of experimental data is unknown or has not been rigorously investigated in a systematic fashion. Hence, the solidity of this bedrock assumption of experimental science lies largely in the realm of belief and trust in the integrity of the authors.

Reproducibility versus replicability. Although many biological scientists intuitively believe that the reproducibility of an experiment means that it can be replicated, Drummond makes a distinction between these two terms (9). Drummond argues that reproducibility requires changes, whereas replicability avoids them (9). In other words, reproducibility refers to a phenomenon that can be predicted to recur even when experimental conditions may vary to some degree. On the other hand, replicability describes the ability to obtain an identical result when an experiment is performed under precisely identical conditions. For biological scientists, this would appear to be an important distinction with everyday implications. For example, consider a lab attempting to reproduce another lab's finding that a certain bacterial gene confers a certain phenotype. Such an experiment might involve making gene-deficient variants, observing the effects of gene deletion on the phenotype, and, if phenotypic changes are apparent, then going further to show that gene complementation restores the original phenotype. Given a high likelihood of microevolution in microbial strains and the possibility that independently synthesized gene disruption and replacement cassettes may have sub-

[▽] Published ahead of print on 27 September 2010.

tly different effects, then the attempt to reproduce findings does not necessarily involve a precise replication of the original experiment. Nevertheless, if the results from both laboratories are concordant, then the experiment is considered to be successfully reproduced, despite the fact that, according to Drummond's distinction, it was never replicated. On the other hand, if the results differ, a myriad of possible explanations must be considered, some of which relate to differences in experimental protocols. Hence, it would seem that scientists are generally interested in the reproducibility of results rather than the precise replication of experimental results. Some variation of conditions is considered desirable because obtaining the same result without absolutely faithful replication of the experimental conditions implies a certain robustness of the original finding. In this example, the replicatibility of the original experiment following the exact protocols initially reported would be important only if all subsequent attempts to reproduce the result were unsuccessful. When findings are so dependent on precise experimental conditions that replicatibility is needed for reproducibility, the result may be idiosyncratic and less important than a phenomenon that can be reproduced by a variety of independent, nonidentical approaches.

Replicability requirement for individual studies. Given the difference between reproducibility and replicability that depends on whether experimental conditions are subject to variation, it is apparent that when most papers state that data are reproducible, they actually mean that the experiment has been replicated. On the other hand, when different laboratories report the confirmation of a phenomenon, it is likely that this reflects reproducibility, since experimental variability between labs is likely to result in some variable(s) being changed. In fact, depending on the number of variables involved, replicability may be achievable only in the original laboratory and possibly by the same experimenter. This accounts for the greater confidence one has in a scientific observation that has been corroborated by independent observers.

The desirability of replicability in experimental science leads to the practical question of how many times an experiment should be replicated before publication. Most reviewers would demand at least one replication, while preferring more. In this situation, the replicability of an experiment provides assurance that the effect is not due to chance alone or an experimental artifact resulting in a one-time event. Ideally, an experiment should be repeated multiple times before it is reported, with the caveat that for some experiments the expense of this approach may be prohibitive. Guidelines for experimentation with vertebrate animals also discourage the use of unnecessary duplication (10, 17). In fact, some institutions may explicitly prohibit the practice of repeating animal experiments that reproduce published results. We agree with the need to repeat experiments but suggest that authors strive for reproducibility instead of simple replicability. For example, consider an experiment in which a particular variable, the level of a specific antibody, is believed to account for a specific experimental outcome, resistance to a microbial pathogen. Passive administration of the immunoglobulin can be used to provide protection and support the hypothesis. Rather than simply replicating this experiment, the investigator might more fruitfully conduct a dose-response experiment to determine the effect of various antibody doses or microbial inocula and test multiple strains rather than simply carrying out multiple replicates of the original experiment.

Limits of replicability and reproducibility. Although the ability of an investigator to confirm an experimental result is essential to good science, with an inherent assumption of reproducibility, we note that there are practical and philosophical limits to the replicability and reproducibility of findings. Although to our knowledge this question has not been formally studied, replicability is likely to be inversely proportional to the number of variables in an experiment. This is all too apparent in clinical studies, leading Ioannidis to conclude that most published research findings are false (13). Statistical analysis and meta-analysis would not be required if biological experiments were precisely replicatable. Initial results from genetic association studies are frequently unconfirmed by follow-up analyses (14), clinical trials based on promising preclinical studies frequently fail (16), and a recent paper reported that only a minority of published microarray results could be repeated (15). Such observations have even led some to question the validity of the requirement for replication in science (21).

Every variable contains a certain degree of error. Since error propagates linearly or nonlinearly depending on the system, one may conclude that the more variables involved, the more errors can be expected, thus reducing the replicability of an experiment. Scientists may attempt to control variables in order to achieve greater reproducibility but must remember that as they do so, they may progressively depart from the heterogeneity of real life. In our hypothetical experiment relating specific antibody to host resistance, errors in antibody concentration, inoculum, and consistency of delivery can conspire to produce different outcomes with each replication attempt. Although these errors may be minimized by good experimental technique, they cannot be eliminated entirely. There are other sources of variation in the experiment that are more difficult to control. For example, mouse groups may differ, despite being matched by genetics, supplier, gender, and age, in such intangible areas as nutrition, stress, circadian rhythm, etc. Similarly, it is very difficult to prepare infectious inocula on different days that closely mirror one another given all the variables that contribute to microbial growth and virulence. To further complicate matters, the outcomes of complex processes such as infection and the host response do not often manifest simple dose-response relationships. Inherent stochasticity in biological processes (19) and anatomic or functional bottlenecks (2) provide additional sources of experiment-to-experiment variability. For many experiments reported in *Infection and Immunity*, the outcome of the experiment is highly dependent on initial experimental conditions, and small variations in the initial variables can lead to chaotic results. In such systems where exact replicability is difficult or impossible to achieve, the goal should be general reproducibility of the overall results. Ironically, results that are replicated too precisely are "too good to be true" and raise suspicions of data falsification (3), illustrating the tacit recognition that biological results inherently exhibit a degree of variation.

To continue the example given above, the conclusion that antibody was protective may be reproduced in subsequent experiments despite the fact that the precise initial result on average survival was never replicated, in the sense that subsequent experiments varied in magnitude of difference observed and time to death for the various groups. Investigators may be able to increase the likelihood that individual experiments are reproducible by enhancing their robustness. A well-known strategy to enhance the likelihood of reproducibility is to increase the power of the experiment by increasing the number of individual measurements, in order to minimize the contri-

bution of errors or random effects. For example, using 10 mice per group in the aforementioned experiment is more likely to lead to reproducible results than using 3 mice, other things being equal. Along the same lines, two experiments using 10 mice each will provide more confidence in the robustness of the results than will a single experiment involving 20 animals, because obtaining similar results on different days lessens the likelihood that a given result was strongly influenced by an unrecognized variable on the particular day of the experiment. When reviewers criticize low power in experimental design, they are essentially worried that the effect of variable uncertainty on low numbers of measurements will adversely influence the reproducibility of the findings. However, subjective judgments based on conflicting values can influence the determination of sample size. For instance, investigators and reviewers are more likely to accept smaller sample sizes in experiments using primates. Consequently, a sample size of 3 might be acceptable in an experiment using chimpanzees while the same sample size might be regarded as unacceptable in a mouse experiment, even if the results in both cases achieve statistical significance. Similarly, cost can be a mitigating factor in determining the minimum number of replicates. For nucleic acid chip hybridization experiments, measurements in triplicate are recommended despite the complexity of such experiments and the range of variation inherent in such measurement, a recommendation that tacitly accepts the prohibitive cost of larger numbers of replicates for most investigators (5). Cost is also a major consideration in replicating transgenic or knockout mouse experiments in which mouse construction may take years. Hence, the power of an experiment can be estimated accurately using statistics, but real-life considerations ranging from the ethics of animal experimentation to monetary expense can influence investigator and reviewer judgment.

We cannot leave the subject of scientific reproducibility without acknowledging that questions about replicability and reproducibility have long been at the heart of philosophical debates about the nature of science and the line of demarcation between science and non-science. While scientists and reviewers demand evidence for the reproducibility of scientific findings, philosophers of science have largely discarded the view that scientific knowledge should meet the criterion that it is verifiable. Through inductive reasoning, Bacon used data to infer that under similar circumstances a result will be repeated and can be used to make generalizations about other related situations (11). However, the logical consistency of such views was challenged by Hume, who posited that inferences from experiences (or, in our case, experiments) cannot be assumed to hold in the future because the future may not necessarily be like the past. In other words, even the daily rising of the sun for millennia does not provide absolute assurance that it will rise the next day. The philosophies of logical positivism and verificationism viewed truth as reflecting the reproducibility of empirical experience, dependent on propositions that could be proven to be true or false. This was challenged by Popper, who suggested that a hypothesis could not be proven, only falsified or not, leaving open the possibility of a rare predictable exception, vividly depicted as the metaphor of a "black swan" (20). One million sightings of white swans cannot prove the hypothesis that all swans are white, but the hypothesis can be falsified by the sight of a single black swan.

A pragmatic approach to reproducibility. Given the challenges of achieving and defining replicatibility and reproducibility in experimental science, what practical guidance can we provide? Despite valid concerns ranging from the true reproducibility of experimental science to the logical inconsistencies identified by philosophers of science, experimental reproducibility remains a standard and accepted criterion for publication. Hence, investigators must strive to obtain information with regard to the reproducibility of their results. That, in turn, raises the question of the number of replications needed for acceptance by the scientific community. The number of times that an experiment is performed should be clearly stated in a manuscript. A new finding should be reproduced at least once and preferably more times. However, even here there is some room for judgment under exceptional circumstances. Consider a trial of a new therapeutic molecule that is expected to produce a certain result in a primate experiment based on known cellular processes. If one were to obtain precisely the predicted result, one might present a compelling argument for accepting the results of the single experiment on moral grounds regarding animal experimentation, especially in situations in which the experiment results in injury or death to the animal. At the other extreme, when an experiment is easily and inexpensively carried out without ethical considerations, then it behooves the investigator to ascertain the replicability and reproducibility of a result as fully as possible. However, there are no hard and fast rules for the number of times that an experiment should be replicated before a manuscript is considered acceptable for publication. In general, the importance of reproducibility increases in proportion to the importance of a result, and experiments that challenge existing beliefs and assumptions will be subjected to greater scrutiny than those fitting within established paradigms.

Given that most experimental results reported in the literature will not be subjected to the test of precise replication unless the results are challenged, it is essential for investigators to make their utmost efforts to place only the most robust data into print, and this almost always involves a careful assessment of the variability inherent in a particular experimental protocol and the provision of information regarding the replicability of the results. In this instance, more is better than less. To ensure that research findings are robust, it is particularly desirable to demonstrate their reproducibility in the face of variations in experimental conditions. Reproducibility remains central to science, even as we recognize the limits of our ability to achieve absolute predictability in the natural world. Then again, ask us next week and you might get a different answer.

REFERENCES

1. **American Physical Society.** 2000. First experiment to draw electricity from lightning, May 10, 1752. APS News **9:**6.
2. **Barnes, P. D., M. A. Bergman, J. Mecsas, and R. R. Isberg.** 2006. *Yersinia pseudotuberculosis* disseminates directly from a replicating bacterial pool in the intestine. J. Exp. Med. **203:**1591–1601.
3. **Bernal, S. K.** 2006. A massive snowball of fraud and deceit. J. Androl. **27:**313–315.
4. **Blount, Z. D., C. Z. Borland, and R. E. Lenski.** 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **105:**7899–7906.
5. **Buck, M. J., and J. D. Lieb.** 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics **83:**349–360.
6. **Casadevall, A., and F. C. Fang.** 2008. Descriptive science. Infect. Immun. **76:**3835–3836.
7. **Casadevall, A., and F. C. Fang.** 2009. Mechanistic science. Infect. Immun. **77:**3517–3519.
8. **Casadevall, A., and F. C. Fang.** 2009. Important science—it's all about the SPIN. Infect. Immun. **77:**4177–4180.
9. **Drummond, C.** 2009. Replicability is not reproducibility: nor is it good science. Proc. Eval. Methods Mach. Learn. Workshop 26th ICML, Montreal, Quebec, Canada. http://www.csi.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf.

10. **Federal Register.** 1989 final rules: animal welfare; 9 CFR parts 1 and 2. Fed. Regist. **54:**36112–36163.
11. **Glass, D. J., and N. Hall.** 2008. A brief history of the hypothesis. Cell **134:**378–381.
12. **Gould, S. J.** Wonderful life: the Burgess Shale and the nature of history. W. W. Norton, New York, NY.
13. **Ioannidis, J. P.** 2005. Why most published research findings are false. PLoS Med. **2:**e124.
14. **Ioannidis, J. P., E. E. Ntzani, T. A. Trikalinos, and D. G. Contopoulos-Ioannidis.** 2001. Replication validity of genetic association studies. Nat. Genet. **29:**306–309.
15. **Ioannidis, J. P., D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello. L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort.** 2009. Repeatability of published microarray gene expression analyses. Nat. Genet. **41:**149–155.
16. **Lowenstein, P. R., and M. G. Castro.** 2009. Uncertainty in the translation of preclinical experiments to clinical trials. Why do most phase III clinical trials fail? Curr. Gene Ther. **9:**368–374.
17. **Nuffield Council on Bioethics.** 2005. The ethics of research involving animals: consensus statement by all members of the working party. Nuffield Council on Bioethics, London, United Kingdom.
18. **Popper, K. R.** 1959. The logic of scientific discovery. Hutchinson, London, United Kingdom.
19. **Shahrezaei, V., and P. S. Swain.** 2008. The stochastic nature of biochemical networks. Curr. Opin. Biotechnol. **19:**369–374.
20. **Taleb, N. N.** 2007. The black swan: the impact of the highly improbable. Random House, New York, NY.
21. **Vieland, V. J.** 2001. The replication requirement. Nat. Genet. **29:**244–245.

Arturo Casadevall
Editor in Chief, mBio
Departments of Microbiology & Immunology and Medicine
Albert Einstein College of Medicine, Bronx, New York

Ferric C. Fang
Editor in Chief, Infection and Immunity
Departments of Laboratory Medicine and Microbiology
University of Washington School of Medicine, Seattle, Washington

---

The views expressed in this Editorial do not necessarily reflect the views of the journal or of ASM.

Editor: A. Camilli