Research Reproducibility, Replicability, Reliability

A Speech by Ralph J. Cicerone, President
National Academy of Sciences
Presented at the Academy's 152nd Annual Meeting
April 27, 2015

Good morning. You are participating in the 152nd annual meeting of the National Academy of Sciences. Last Saturday, we welcomed the NAS members and foreign associates who were elected in 2014, individuals with great achievements in science. A representative group of them also spoke on Saturday and gave extremely interesting descriptions of discoveries and of prospects in differing fields of science. On Sunday, NAS awards and prizes were presented for very notable work across many fields of science and our Public Welfare Medalist, Neil deGrasse Tyson spoke about key aspects of NAS's mission. And the Foucault pendulum is hanging once again in the Great Hall!

Reproducibility in Research

Science depends on curiosity, inspiration, observations, formulations, calculations, resources, communication of ideas and on reproducibility.

The reproducibility of scientific experiments and calculations embodies a fundamental aspect of science. As we all know, the ability to replicate experiments and to verify findings by using independent methods, or by new investigators simply repeating earlier experiments, is critical to the advancement of science. This ability can be used to formally test hypotheses or it can test particular predictions or simply confirm observations of system variables, for example, in response to the manipulation of a variable.

Efforts to reproduce experiments enable science to be self-correcting, and this quality of science is very special and rare amongst human endeavors.

We also know that independent verification of experimental results can guide subsequent research, for example, by solidifying previous findings. When previous results are not confirmed or strengthened, the new results can allow new ideas and hypotheses to be generated, and avoid wasting of time and wild goose chases. In this pursuit, it helps greatly to describe an experiment so that it can be repeated.

Indeed, it is essential to describe in written publications, experimental methods, materials and results clearly so that subsequent research can be compared. Sometimes follow-up experiments occur soon, sometimes years later, and often with very different methods and instruments. In all cases, the written record is essential as is access to experimental data. This process enables science to be self-correcting.

I don't know of any better way!

Thus, the reporting of research results itself is clearly an essential step in science's progress. If you don't report results, no one can know of what you have done. It is as if you have done nothing. Yet, complications arise in reporting that can impede the completeness of reporting of methods, initial conditions, assumptions, characteristics of data, and statistical methods. For example, today there is emphasis on brevity and speed of publication. Increasing numbers of publications and of journals, some with differing standards and modes of operation, also can impede full reporting of experimental conditions and results. Additionally, various bodies now employ metrics of publication

Figure 1 © The Economist Newspaper Limited, London (Oct 2013).

as they evaluate individuals, programs, fields and institutions, and some metrics can act as disincentives, against rather than encourage reproducibility in research.

Today, I want to increase awareness of some signs of trouble in the reproducibility of research and to identify some of the topics in which we as researchers can help to improve prospects for scientific progress and for our institutions while we also increase public respect and support for science as an enterprise. A lot is going on at this front.

Signs of Problems

Each of us can recall some instances of erroneous or irreproducible research that have become high-profile cases (e.g., a Lancet paper on vaccines; stem-cell claims from researchers in several countries; a large problem some years ago in materials physics). I will not elaborate any further about specific incidents but there certainly have been cases that have angered independent experts and that have aroused public attention and led to institutional investigations. In some cases, large claims have been made that were shown to be erroneous or even fraudulent. While these kinds of cases are indeed serious, they also provide evidence that science *is* self-correcting, that is, that mistakes and/or misconduct are discovered and are addressed. In fact, when large claims are made on topics of great commercial, medical, regulatory or environmental interest, competing scientists often mount experiments quickly to test the original claims, as happened with claims of cold fusion.

However, broader reports of irreproducible research have been offered recently by individual scientists and by mainstream newspapers. (*The Economist, New York Times, Washington Post, The Guardian*…). For example, *The Economist* article, "Trouble at the Lab", (1) notes problems and practices in a variety of scientific fields and it quotes scientists who have become aware of such problems [Figure 1]. The article goes on to examine practices that might be contributing to errors, from challenges of experimental design to statistical analysis and sloppiness under pressure for results, and more. Despite the rather lurid cartoon [Figure 2] *The Economist* article was a fair one.

Disturbing indicators of the frequency of research that cannot be replicated are also spotlighted in *The Economist* article, such as a "Comment" to *Nature* in which Amgen researchers (2) reported that their attempts to confirm results from 53 papers in preclinical cancer research succeeded in only six cases. The Amgen researchers cited a similar published study (3) from Bayer Health Care (Germany) in which only 25% of similar studies could be confirmed well enough to justify continuation of research and development. [Figure 3]

The Economist reported that some experiments in psychology, on priming of decisions, have led to erroneous claims, while research in machine learning has displayed some flaws, and in other fields, some notable submissions of intentionally flawed phony papers have been accepted in many journals. There are more such examples in other fields.

Figure 2 ©*Jason Ford/Heart Agency*

How seriously should these reports be taken? Many questions arise about the frequency of incidents and whether they are increasing. How damaging are they to science itself? How much is statistical analysis involved? How much misconduct or fraud is there? Are some experiments not replicating earlier results compromised by the use of poor materials or of poorly characterized experimental conditions and materials? Are inherent complexities of experiments, e.g., those on living organisms or cells, partly to blame? Is the haste to publish and to publish often going greatly awry? Is the use of extremely large data sets (too large for anyone to be familiar with all aspects of them) leading to confusion and errors? Are multi-authored and multi-discipline papers more prone to slip-ups? Are reviewers and editors overloaded? What should be done and by whom?

Responses, Reproducibility Initiatives …

Let us consider what kinds of responses are already underway or are called for as scientists seek even more powerful methods of self-correction. In some specific fields of science, projects are being mounted that are squarely aimed at determining causes of difficulty in replicating findings and at improving capabilities. The "Reproducibility Project: Cancer Biology" (4) aims to conduct independent experiments to determine if findings from 50 prominent experiments (in pre-clinical cancer biology research) can be replicated. Those 50 experiments with high potential impact were identified from published literature of 2010-2012. In an ongoing experiment, their authors are being contacted and replication experiments are being conducted by researchers in third-party laboratories coordinated by the Science Exchange and the Center for Open Science. This project is funded by the Laura and John Arnold Foundation.

The "Reproducibility Project: Psychology" (5) is aimed at determining how replicable are key results from selected journal publications. The project focuses on journal articles from three selected psychological journals and scientists from all over the world volunteer to replicate a study of their choosing from those journals.

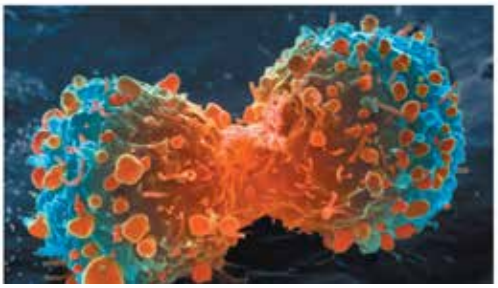

Figure 3 *Reprinted by permission from Macmillan Publishers LTD: Nature 483, 531-33 (29 March 2012).*

Raise Standards for Preclinical Cancer Research. C. Glenn Begley and Lee M. Ellis propose how methods, publications, and incentives must change if patients are to benefit.

Experimenters agree to follow a structured protocol for designing and conducting potential replications of key published findings. Open source software facilitates open collaboration in this research. The project is overseen by the Center for Open Science (University of Virginia) and it is funded by the Laura and John Arnold Foundation. Efforts to deal productively with issues of replicability in psychological "priming" research are also underway (6).

Both of these reproducibility projects are field-specific (cancer biology and psychology). There is good reason for this approach even though there are some counter arguments. For example, our NAS/NRC report (2009) "Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age" concluded that attention and actions from specific fields of science are needed to achieve what the report title describes — ensuring the integrity, accessibility and stewardship of research data — partly because data from different fields differ in how they are collected, analyzed and stored. [Figure 4]

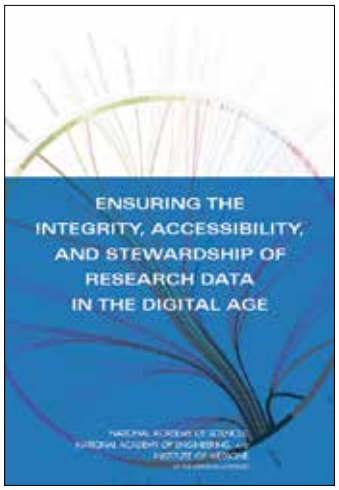

Figure 4

While principles of data sharing and publication strongly favor openness and transparency there are some important limitations that arise field-by field: personal privacy, national security, proprietary and commercial interests and agreements and the use of data in government regulations, for example. Each of these limitations can also impede how tests of replicability can be conducted (other terms are used, including "reproducibility" and "reliability").

NAS member Gary King launched a political science-specific replicability project about twenty years ago and it continues (7). The project "Replication Replication" is aimed at improving the norms of data sharing and replication in scholarly political science and quantitative social science research. King has called for fundamentally new standards, protocols, and software for citing, sharing, analyzing, archiving, preserving, distributing, cataloging, translating, disseminating, naming, verifying, and replicating research data and analyses (7). King also launched a project "Publication Publication" aimed at students and how to conduct interesting research involving replicability, and a third project DataVerse that has established data repositories as a resource to researchers in several fields of science. The DataVerse Network project (8) promotes sharing, citing, using, and archiving scientific data for reproducible research.

Different fields of science may have differing views of how seriously questions of reproducibility should be taken and how to deal with them. The NAS/NRC is undertaking efforts in individual fields to clarify questions and issues. For example, the NRC's Division on Behavioral and Social Sciences and Education is beginning a project on research replicability in the behavioral and social sciences (sponsored by the Arnold Foundation). Also, the NRC Committee on Applied and Theoretical Statistics (CATS), (a standing committee of the Board on Mathematical Sciences and their Applications) recently convened a workshop to explore whether improved statistical methods, or improved application of existing methods, could increase the reproducibility of research. They asked questions including: how can we quantify the reproducibility of scientific results? Because variability across studies is to be expected, how can we assess the acceptable degree of variability and when should we be concerned about reproducibility? How can the choice of statistical methods for designing a study and analyzing its output data affect the reproducibility of a scientific result? Are there analytical approaches that can enhance reproducibility, within disciplines and overall? Do we need new conceptual/theoretical frameworks for assessing the strength of evidence from a study?

Specific actions are being taken in some quarters to improve the statistical treatment of research data. In one such endeavor, *Science* magazine Chief Editor Marcia McNutt has established (9) a Statistical Board of Reviewing Editors (SBoRE).

This group of experts will contribute to the review of manuscripts submitted to *Science* and will evaluate data-analysis methods and their application to interpretations, to measures of statistical uncertainty, and therefore to reproducibility.

Broader issues surrounding research reproducibility are also being considered by journals. For example, a group of editors from over 30 journals met last year to discuss principles and guidelines to increase transparency and reproducibility in preclinical biomedical research (10). Principles, some of which have been adopted by certain journals, focus on statistical methods, experimental design and details, accessibility to data and to experimental materials. Guidelines deal with best practices for image-based data and experimental description including fuller specifications on laboratory animals and cell lines.

Also, *eLife* has arranged to review some Cancer Biology project proposals, or Registered Reports, for many of the top 50 Cancer Biology papers (that were selected by the Center for Open Science and the Science Exchange), and have published several of them (11). The groups whose reports were published are now engaged in efforts to reproduce key experiments from these papers that were judged appropriate by members of the *eLife* Board and *ad hoc* referees.

This variety of actions is occurring in response to identified cases of non-replicable research or in efforts to continuously improve how science functions. But there are other potential reasons to improve, notably to head off top-down governmental edicts. Recently, one committee of Congress is discussing a call for NSF to support an NAS/NRC project "to assess research and data reproducibility and replicability issues in interdisciplinary research and to make recommendations on how to improve rigor and transparency in scientific research."

Conducting, Reporting and Evaluating Research

Being an active scientific researcher is a privilege. Making discoveries is extremely rewarding and fascinating. Watching and understanding the advances made by others is also very stimulating, as is sensing their potential applications. Frontiers of understanding are being extended in ways that were unforeseen in earlier ages, and additional progress is easy to imagine. Scientific competition is a powerful force that helps to drive this progress.

Moreover, scientific research enables better teaching too, for example, by involving undergraduate and graduate students along with postdoctoral fellows and senior researchers where they can emphasize scientific ways of thinking. Educational benefit also arises by inviting school teachers into research projects at universities and government laboratories.

Nonetheless, there are also patterns in research today that cause concern and we must address the ones over which scientists have some control. The overall growth of science worldwide, a positive development, has also led to many more journals of variable quality and to more disciplines with differing practices. The complexity of many scientific questions has led to the creation of large research groups and the need for support staff and more equipment. Competition for research support has become more strained and there is perception of hypercompetition that leads to deterioration in how science is conducted, for example, in biomedical research (12).

The reporting of research results, essential as it is, has become burdened by many excesses. Generally, there is pressure to publish more often and to publish briefer papers quickly. The need to demonstrate impact of one's research is often manifested by citing numbers of papers. Large collaborations with multiple authors (often in other countries) make it more difficult to report results in ways that encourage replication and independent verification. The desirability of publishing in the most selective journals (which favor papers that report something really new and different) has simultaneously diminished the rewards of confirming and/or testing previously published papers.

What happens when errors are published? Ideally, journals publish retractions of individual papers. In cases where researchers find their own mistakes and publish a retraction (or a *self-corrigendum*), science is well served and the researchers are respected. Science also benefits when other researchers publish independent corrections. Yet there are cases where authors refuse to retract (even when co-authors are willing) for which neither the journals nor the researcher's home institution have the ability to conduct inquiries. Indeed, the embarrassment and stigma of retractions are disincentives for institutions to pursue inquiries as to retractions or even to investigate misconduct. Earlier this year, at a National Academy of Sciences retreat (at the Annenberg Foundation's Sunnylands) "Ensuring the Integrity of Science" (13), it was suggested that there is need for two different kinds of paper retractions: "voluntary" and "for cause". The former category would be seen as a positive contribution and the latter would be viewed negatively.

In most fields, we should sort out rules for co-authorship of research publications. Guidelines on how to decide co-authorship vary greatly by discipline. At one time in my career, I thought that every co-author should be able to describe everything in the paper. Now, large data sets gathered by specialized instruments and or complex calculations can be provided by participants who are necessary but who might not be familiar with all research methods and results. Being generous with coauthorship is a good rule

but it can compromise efforts to report so as to facilitate replicability. Authorship as a courtesy to senior investigators can also be discouraging to other contributors and can impede replicability. The many issues of authorship and of credit and responsibility for publications need re-examination, probably inside individual research disciplines.

Evaluations of individual researchers, as in hiring and career advancement are highly important functions and involve issues related to publications. Many individuals have noticed that citation data for individuals and for journals themselves are being used indiscriminately, giving at least the appearance of replacing more fundamental measures of quality. Selected citation data are useful but we must choose them carefully. While workloads of editors, reviewers and of academic advancement committees are heavy, we must insist that evaluators read papers to evaluate quality and reject cavalier citation data (often generated automatically by machines with lots of decimal points) (14, San Francisco Declaration on Research Assessment). Evaluation of individual researchers is an example of a function over which researchers have much control, yet one faculty member at a leading research university complained to me that his/her dean was in charge and that "he wants citations". Deans who overuse citation data can be replaced by faculties.

Signs of progress in this direction can be seen in some evaluation processes that limit the number of publications that may be considered or submitted, for example, nomination for NAS membership, funding proposals to the National Science Foundation and the National Institutes of Health and academic advancement procedures at some research universities. Similar progress can be achieved elsewhere such as the evaluation of research programs and of institutions.

What's Next?

In preparing this summary of questions about replicability of research and of ongoing efforts to improve prospects for replicability, I have learned that there is much good work underway. Yet there is much to do. The academy itself should help in some key ways. For example, we can convene groups to learn what is working in individual fields and journals. We can also provide useful guidelines toward best practices for data sharing and access to experimental materials such as PNAS's current instructions to authors (15).

I am very encouraged by reproducibility experiments that I have learned about, like those in cancer biology and in psychology and the extended projects of Professor Gary King, and efforts on data sharing on statistics and the actions of several journals. Let's continue these experiments and evaluate their results.

There are roles in these efforts for individual researchers, university and government-laboratory leaders, journal editors and reviewers and disciplinary scientific societies and wherever career advancement and research awards are decided. There are also major incentives for us all: strengthening the ability of science and individual scientists to make progress and to continue to command respect from the general public. I hope that each of us will pitch in to help to assure the self-correcting aspect of science through reproducibility of research and access to relevant data. The important related issues of research publications and of evaluation of research also need our continuing attention.

In all of these topics, there are talented and interested journalists who cover science who are helping to sense public interests and who can aid many scientists to address questions of the significance of specific research results along with broader issues. We should work with them when given the opportunity, to show how we are improving.

Separately, tomorrow, Tuesday April 28 there will be a breakout session here in our Lecture Room led by Randy Schekman focused exactly on these topics.

At NAS we will continue to improve our efforts toward self-correction in science and we will issue reports as we proceed (upcoming report from COSEPUP, from meetings focused on

replicability and on data access) and a brief summary of the NAS-Annenberg Foundation retreat.

I look forward to seeing more encouraging progress toward sharpening the ability of science to be self-correcting.

LINKS AND REFERENCES

(1) *Economist*, October 19-25 (2013), "Trouble at the Lab". [http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble]

(2) C. Glenn Begley and Lee M. Ellis, "Raise standards for preclinical cancer research", *Nature* 483, 531-533 (2012). [http://www.nature.com/nature/journal/v483/n7391/full/483531a.html]

(3) Prinz, F., Schlange, T. and Asadullah, K., *Nature* Rev. Drug Discov. 10, 712 (2011). [http://www.nature.com/nrd/journal/v10/n9/full/nrd3439-c1.html]

(4) Reproducibility Project: Cancer [https://osf.io/e81xl/]

(5) Reproducibility Project: Psychology [https://osf.io/ezcuj/]

(6) An early suggestion was offered by NAS member D. Kahneman. [http://www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf]

(7) Gary King, "Replication Replication" [http://gking.harvard.edu/files/replication.pdf]

(8) Dataverse network [https://thedata.harvard.edu/dvn/]

(9) "Raising the Bar" M. McNutt, *Science* 345, p. 9 (2014) [http://www.sciencemag.org/content/345/6192/9.full]

(10) Journals unite for reproducibility" M. McNutt, *Science* 346, p. 679 (2014) [http://www.sciencemag.org/content/346/6210/679.full]

(11) eLife [http://elifesciences.org/collections/reproducibility-project-cancer-biology]

(12) as described by Alberts et. al. (PNAS, 2014). Whether the solutions proposed here are ideal is still being sorted out. [http://www.pnas.org/content/111/16/5773.full]

(13) National Academy of Sciences Retreat "On Ensuring the Integrity of Science" at Annenberg Foundation's Sunnylands, February, 2015.

Participants were: Ralph Cicerone, Bruce Alberts, Stephen E. Fienberg, Alexander Kamb, Marcia McNutt, Robert Nerem, Randy Schekman, Richard Shiffrin, Victoria Stodden, Subra Suresh, Maria Zuber, Kathleen Hall Jamieson and Barbara Kline Pope. A brief summary report from this group is planned.

(14) San Francisco Declaration on Research Assessment [http://www.ascb.org/dora-old/files/SFDeclarationFINAL.pdf]

(15) In PNAS Instructions for authors, see "Data Sharing" under "Journal Policies." [http://www.pnas.org/site/authors/format.xhtml]