PERSPECTIVE

How measurement science can improve confidence in research results

Anne L. Plant[1]*, Chandler A. Becker[1], Robert J. Hanisch[1], Ronald F. Boisvert[2], Antonio M. Possolo[2], John T. Elliott[1]

1 Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America, **2** Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, United States of America

* anne.plant@nist.gov

Abstract

The current push for rigor and reproducibility is driven by a desire for confidence in research results. Here, we suggest a framework for a systematic process, based on consensus principles of measurement science, to guide researchers and reviewers in assessing, documenting, and mitigating the sources of uncertainty in a study. All study results have associated ambiguities that are not always clarified by simply establishing reproducibility. By explicitly considering sources of uncertainty, noting aspects of the experimental system that are difficult to characterize quantitatively, and proposing alternative interpretations, the researcher provides information that enhances comparability and reproducibility.

OPEN ACCESS

Citation: Plant AL, Becker CA, Hanisch RJ, Boisvert RF, Possolo AM, Elliott JT (2018) How measurement science can improve confidence in research results. PLoS Biol 16(4): e2004299. https://doi.org/10.1371/journal.pbio.2004299

Published: April 23, 2018

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: BIPM, International Bureau of Weights and Measures (Bureau International des Poids et Mesures); GRADE, Grades of Recommendation, Assessment, Development and Evaluation; GUM, *Guide to Expression of Uncertainty in Measurement*; IEC, International Electrotechnical Commission; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; ILAC, International Laboratory Accreditation Cooperation; ISO, International Organization for Standardization; IUPAP, International Union of Pure and Applied Physics; MIACA, Minimum Information About a Cellular

Indicators of confidence in research results

While reports about the difficulty of reproducing published biomedical research results in the labs of pharmaceutical companies [1,2] have in large part triggered the current "reproducibility crisis," reproducibility has also been cited as a concern in computation [3], forensics [4], epidemiology [5], psychology [6], and other fields, including chemistry, biology, physics and engineering, medicine, and earth and environmental sciences [7].

While "reproducibility" is the term most often used to describe the issue, it has been frequently pointed out that reproducibility does not guarantee that a result of scientific inquiry tracks the truth [8–11]. It has been suggested that, instead, there is a need for "a fundamental embrace of good scientific methodology" [12], and the term "metascience" has been proposed to refer to the idea that rigorous methods can be used to examine the reliability of results [13].

These perspectives suggest that it would be worthwhile to consider how the concepts of measurement science—i.e., metrology—can provide useful guidance that would enable researchers to assess and achieve rigor of a research study [14]. The goal of measurement science is comparability, which enables evaluation of the results from one time and place relative to results from another time and place; this is ultimately the goal of establishing rigor and reproducibility. The purpose of this manuscript is to provide a practical connection between the field of metrology and the desire for rigor and reproducibility in scientific studies.

Assay; MIAME, Minimum Information about a Microarray Experiment; MIATA, Minimum Information About T-cell Assays; NIH, National Institutes of Health; NIST, National Institute of Standards and Technology; RDA, Research Data Alliance; VIM, *International Vocabulary of Basic and General Terms in Metrology*.

Provenance: Not commissioned; externally peer reviewed.

In the field of metrology, a measurement consists of two components: a value determined for the measurand and the uncertainty in that value [15]. The uncertainty around a value is an essential component of a measurement. In the simplest case, the uncertainty is determined by the variability in replicate measurements, but for complicated measurements, it is estimated by the combination of the uncertainties at every step in the process. The concepts that support quantifying measurement uncertainty arise from international conventions that have been agreed to through consensus by scientists in many fields of study over the past 150 years and continue to be developed. These conventions are developed and adopted by the National Metrology Institutes around the world (including the National Institute of Standards and Technology [NIST] in the United States) and international standards organizations such as the International Bureau of Weights and Measures (Bureau International des Poids et Mesures, BIPM), the International Electrotechnical Commission (IEC), the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC), the International Organization for Standardization (ISO), the International Union of Pure and Applied Physics (IUPAP), the International Laboratory Accreditation Cooperation (ILAC), and others. These efforts helped to advance the concepts of modern physics by providing the basis on which comparison of data was made possible [14]. Thus, it seems appropriate to examine these concepts today to inform our current concerns about rigor and reproducibility.

One of the consensus documents developed by measurement scientists is the *Guide to Expression of Uncertainty in Measurement* [16], commonly known as the GUM. This document describes the types of uncertainty (e.g., Type A, those that are evaluated by statistical methods; and Type B, those that are evaluated by other means) and methods for evaluating and expressing uncertainties. The GUM describes a rigorous approach to quantifying measurement uncertainty that is more readily applied to well-defined physical quantities with discrete values and uncertainties (such as the measurements of amount of a substance, like lead in water) than to measurements that involve many parameters (such as complex experimental studies involving cells and animals). Calculating uncertainties in such complex measurement systems is a topic of ongoing research. But even if uncertainties are not rigorously quantified, the concepts of measurement uncertainty provide a systematic thought process about to how to critically evaluate comparability between results produced in different laboratories.

The GUM identifies examples of sources of uncertainty. These include an incomplete definition of what is being measured (i.e., the measurand); the possibility of nonrepresentative or incomplete sampling, in which the samples measured may not represent all of what was intended to be measured; the approximations and assumptions that are incorporated in the measurement method and procedure; and inadequate knowledge of the effects of environmental conditions on the measurement. In Table 1, we have grouped the sources of uncertainty identified in the GUM that are common to many scientific studies, and we have indicated measurement science approaches for characterizing and mitigating uncertainty.

The GUM also provides definitions of many terms such as “repeatability” (which is defined as the closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement) and “reproducibility” (which is defined as the closeness of the agreement between the results of measurements of the same measurand carried out under different conditions of measurement). A complete list of consensus definitions of measurement-related terms can be found in the *International Vocabulary of Basic and General Terms in Metrology* (VIM) [18]. A recent publication demonstrates the adoption of these definitions to harmonize practices across the geophysics community [19].

Table 1. Identifying, reporting, and mitigating sources of uncertainty in a research study.

1. State the plan

a. Clearly articulate the goals of the study and the basis for generalizability to other settings, species, conditions, etc., if claimed in the conclusions.

b. State the experimental design, including variables to be tested, numbers of samples, statistical models to be used, how sampling is performed, etc.

c. Provide preliminary data or evaluations that support the selection of protocols and statistical models.

d. Identify and evaluate assumptions related to anticipated experiments, theories, and methods for analyzing results.

2. Look for systemic sources of bias and uncertainty

a. Characterize reagents and control samples (e.g., composition, purity, activity, etc.).

b. Ensure that experimental equipment is responding correctly (e.g., through use of calibration materials and verification of vendor specifications).

c. Show that positive and negative control samples are appropriate in composition, sensitivity, and other characteristics to be meaningful indictors of the variables being tested.

d. Evaluate the experimental environment (e.g., laboratory conditions such as temperature and temperature fluctuations, humidity, vibration, electronic noise, etc.).

3. Characterize the quality and robustness of experimental data and protocols

a. Acquire supplementary data that provide indicators of the quality of experimental data. These indicators include precision (i.e., repeatability, with statistics such as standard deviation and variance), accuracy (which can be assessed by applying alternative [orthogonal] methods or by comparison to a reference material), sensitivity to environmental or experimental perturbants (by testing for assay robustness to putatively insignificant experimental protocol changes), and the dynamic range and response function of the experimental protocol or assay (and assuring that data points are within that valid range).

b. Reproduce the data using different technicians, laboratories, instruments, methods, etc. (i.e., meet the conditions for reproducibility as defined in the VIM).

4. Minimize bias in data reduction and interpretation of results

a. Justify the basis for the selected statistical analyses.

b. Quantify the combined uncertainties of the values measured using methods in the GUM [16] and other sources [17].

c. Evaluate the robustness and accuracy of algorithms, code, software, and analytical models to be used in analysis of data (e.g., by testing against reference datasets).

d. Compare data and results with previous data and results (yours and others').

e. Identify other uncontrolled potential sources of bias or uncertainty in the data.

f. Consider feasible alternative interpretations of the data.

g. Evaluate the predictive power of models used.

5. Minimize confusion and uncertainty in reporting and dissemination

a. Make available all supplementary material that fully describes the experiment/simulation and its analysis.

b. Release well-documented data and code used in the study.

c. Collect and archive metadata that provide documentation related to process details, reagents, and other variables; include with numerical data as part of the dataset.

Abbreviations: GUM, *Guide to Expression of Uncertainty in Measurement*; VIM, *International Vocabulary of Basic and General Terms in Metrology*

https://doi.org/10.1371/journal.pbio.2004299.t001

What does Table 1 add to existing efforts?

There have been many efforts to encourage more reliable research results, and many fields have proposed or instituted conventions, checklists, requirements, and reporting standards that are applicable to their specific disciplines. Some of these include the Grades of Recommendation, Assessment, Development and Evaluation (GRADE) approach for assessing clinical evidence [20], the minimum information activities that have a long history in the biosciences (e.g., Minimum Information about a Microarray Experiment [MIAME]) [21], checklists developed by scientific journals requiring specific criteria to be reported [22], a NIST system for checking thermodynamic data prior to publication [23], and many more. These efforts are not intended to be comprehensive determinations of potential sources of uncertainty in measurement. But interest in measurement science principles is increasing. For example, the Minimum Information About a Cellular Assay (MIACA) activity [24], which was last updated in 2013, encourages reporting the experimental details of cellular assay projects. The more recent Minimum Information About T cell Assays (MIATA), [25,26] which is focused on identifying and encouraging the reporting of variables of particular importance to the outcome of T cell assays, is more comprehensive. MIATA guidelines go beyond descriptions of activities and reagents to include the reporting of quality control activities such as providing information regarding the strategies for data analysis and reporting any effort to pretest medium or serum for assay performance. The most current National Institutes of Health (NIH) instructions for grant applications [27] speak to many of the concepts of metrology: stating the scientific premise and considering the strengths and weaknesses of prior research; applying scientific method to experimental design, methodology, analysis, and interpretation; considering biological variables such as sex; and authenticating biological and chemical resources that may be sources of variability. Thus, it seems timely to suggest a comprehensive framework that can help to guide identification of the many other potential sources of uncertainty. The conceptual framework in Table 1 can enhance existing guidelines by helping scientists identify potential sources of uncertainty that might not have been considered in existing checklists and to provide some strategies for reducing uncertainty. Table 1 is designed to help guide researchers' critical thinking about the various aspects of their research in an organized way that encourages them to document the data they can, and often do, collect that provide confidence in the results.

The inclusion of supporting evidence helps end users of research results—such as decision-makers, commercial developers, and other researchers—know how best to use and follow up on the results. Few research studies will address all aspects indicated in Table 1. But by explicitly acknowledging what is known—or, more importantly, what isn't known—about the various components of a research effort, it is easier to see the strengths and limitations of a study and to assess, for example, whether the study is more preliminary in nature or if the results are highly reliable. The Data Readiness Level is a concept that has been put forward by the nanotechnology community and is an example of this kind of approach, [28] and others have suggested the need for this level of reporting [11].

What are the hurdles that keep ideas such as these from being implemented?

The sociological issues that accompany the "reproducibility crisis" have been discussed in many venues and are beyond the scope of this discussion. Instead, we focus on the principles and practices of measurement science since we find that researchers, particularly in rapidly advancing fields, are sometimes confused about how to apply these principles of the scientific method to achieve "rigor and reproducibility."

A hurdle to implementation of these concepts is the need for tools and technologies that can reduce the challenges for experimentalists who want to address the elements in Table 1. There has not been sufficient investment, perhaps, in technologies that could allow us to better characterize the components of our experimental systems, such as antibody reagents, cell lines, or image analysis pipelines. As a scientific community, we have not prioritized investments in software to facilitate collecting information on complex experimental protocols. While there is great interest in data mining, there is still a lack of progress in the development of natural language and other approaches for achieving harmonized vocabularies that would make it easier to compare and share experimental metadata and protocols. Efforts associated with capturing the details of complicated experimental protocols are being undertaken. PLOS has entered into a collaboration with Protocols.io [29] to facilitate reporting, sharing, and improving protocols. Another effort, ProtocolNavigator [30], enables collection of highly detailed experimental information and storage of provenance information; there are also supporting links to stored data and explanatory videos [31]. Challenges associated with data and digital resources are being considered by the Research Data Alliance (RDA) [32]. The RDA was established in 2013 to foster the sharing of research data but recognized that effective sharing requires standards and best practices and is pursuing technical developments in data discovery, semantics, ontologies, data citation and versioning, data types, and persistent identifiers. Also, with the current emphasis on open data [33] and large-scale data sharing [32], it would be helpful to have a means of evaluating the aspects of the research that establish confidence of the results being shared, especially by those who are using data outside of their area of technical expertise. In addition, increased support for the science that underpins the technologies and methods that help to establish confidence in data will contribute to improving the reusability of published research results.

Conclusions

The consideration by researchers of a systematic approach to identifying sources of uncertainty will enhance comparability of results between laboratories. Because no single scientific observation reveals the absolute "truth," the job of the researcher and the reviewer is to determine how ambiguities have been reduced and what ambiguities still exist. By addressing and characterizing the components of the study as potential sources of uncertainty, the researcher can provide the supporting evidence that helps to define the characteristics of the data, analysis, and tests of the assumptions that were made; such evidence provides confidence in the results and helps inform the reader about how to use the information. Unfortunately, even when studies include these activities, they are rarely reported in an explicit and systematic way that provides maximum value to the reader.

A framework such as the one outlined in Table 1 is applicable to many areas of scientific research. The ideas presented here are not radical or new but are worthy of reconsideration because of the current concern about comparability of research results. We provide this information in the spirit of stimulating discussion within and among the scientific disciplines. More explicit use and documentation of the concepts discussed above will improve confidence in published research results. Applying these concepts will require commitment and critical thinking on the part of individuals, as well as a continuation of the tradition of cooperative effort within and across scientific communities. The end result will be worth the additional effort.

References

1. Begley C. G. & Ellis L. M. Drug development: Raise standards for preclinical cancer research. *Nature* 2012; 483(7391): 531–533 https://doi.org/10.1038/483531a PMID: 22460880
2. Prinz F., Schlange T. & Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 2011; 10(9): 712, https://doi.org/10.1038/nrd3439-c1 PMID: 21892149
3. Peng R. D. Reproducible research in computational science. *Science* 2011; 334(6060): 1226–1227, https://doi.org/10.1126/science.1213847 PMID: 22144613
4. Strengthening Forensic Science in the United States: A Path Forward. Report No. 978-0-309-13130-8 (Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, 2009).

5. Ioannidis J. P. et al. A network of investigator networks in human genome epidemiology. *Am J Epidemiol* 2005; 162(4): 302–304, https://doi.org/10.1093/aje/kwi201 PMID: 16014777
6. Open Science, C. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 2015; 349(6251): aac4716, https://doi.org/10.1126/science.aac4716 PMID: 26315443
7. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016; 533: 452–454, https://doi.org/10.1038/533452a PMID: 27225100
8. Roush S. *Tracking Truth*: *Knowledge*, *Evidence and Science*. (Oxford: Oxford University Press, 2006).
9. Leek J. T. & Peng R. D. Opinion: Reproducible research can still be wrong: adopting a prevention approach. *Proc Natl Acad Sci U S A* 2015; 112: 1645–1646, https://doi.org/10.1073/pnas.1421412111 PMID: 25670866
10. Plant A. L., Locascio L. E., May W. E. & Gallagher P. D. Improved reproducibility by assuring confidence in measurements in biomedical research. *Nat Methods* 2014; 11: 895–898, https://doi.org/10.1038/nmeth.3076 PMID: 25166868
11. Goodman S. N., Fanelli D. & Ioannidis J. P. What does research reproducibility mean? *Sci Transl Med* 2016; 8: 341ps312, https://doi.org/10.1126/scitranslmed.aaf5027 PMID: 27252173
12. Baker M. US societies push back against NIH reproducibility guidelines. *Nature* 2015, https://doi.org/10.1038/nature.2015.17354
13. Schooler J. W. Metascience could rescue the 'replication crisis'. *Nature* 2014; 515, 9, https://doi.org/10.1038/515009a PMID: 25373639
14. Sene M., Gilmore I. & Janssen J. T. Metrology is key to reproducing results. *Nature* 2017; 547: 397–399, https://doi.org/10.1038/547397a PMID: 28748943
15. Possolo A. & Iyer H. K. Invited Article: Concepts and tools for the evaluation of measurement uncertainty. *Rev Sci Instrum* 2017; 88: 011301, https://doi.org/10.1063/1.4974274 PMID: 28147677
16. Evaluation of measurement data—Guide to the expression of uncertainty in measurement. Report No. JCGM 100:2008, (BIPM, 2008).
17. International Vocabulary of Metrology—Basic and General Concepts and Associated Terms. Report No. JCGM 200:2012, (BIPM, 2012).
18. Loew A. et al. Validation practices for satellite-based Earth observation data across communities. *Rev Geophys* 2017; 55: 779–817, https://doi.org/10.1002/2017rg000562
19. Box G. E. P., Hunter J. Stuart, Hunter William G. *Statistics for experimenters*: *design*, *discovery and innovation*. (Hoboken, NJ: John Wiley and Sons, Inc., 2005).
20. Atkins D. et al. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328: 1490, https://doi.org/10.1136/bmj.328.7454.1490 PMID: 15205295
21. Brazma A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat.Genet*. 2001; 29: 365–371 https://doi.org/10.1038/ng1201-365 PMID: 11726920
22. Enhancing Reproducibility. *Nat.Methods* 2013; 10: 367 PMID: 23762900
23. Frenkel M. et al. ThermoData engine (TDE): Software implementation of the dynamic data evaluation concept. *J Chem Inf Model* 2005; 45: 816–838, https://doi.org/10.1021/ci050067b PMID: 16045275
24. Minimum Information About a Cellular Assay [Internet], 2013 Mar 27 [cited 4 April 2018]. https://sourceforge.net/projects/miaca/
25. Britten C. M. et al. T cell assays and MIATA: the essential minimum for maximum impact. *Immunity* 2012; 37: 1–2, https://doi.org/10.1016/j.immuni.2012.07.010 PMID: 22840835
26. Minimum information about a T-cell assay [Internet], [cited 4 April 2018]. http://miataproject.org/implementation/tool-checklist/
27. Updated Application Instructions to Enhance Rigor and Reproducibility. 2017 Dec 13 [cited 4 Apr 2018. https://www.nih.gov/research-training/rigor-reproducibility/updated-application-instructions-enhance-rigor-reproducibility.
28. Nanotechnology Signature Initiative: Nanotechnology Knowledge Infrastructure (NKI) Data Readiness Levels discusssion draft. (2013).
29. Protocols.io Tools for PLOS Authors: Reproducibility and Recognition [Internet] http://blogs.plos.org/plos/2017/04/protocols-io-tools-for-reproducibility/ (2017).
30. Khan I. A. et al. ProtocolNavigator: emulation-based software for the design, documentation and reproduction biological experiments. *Bioinformatics* 2014; 30: 3440–3442, https://doi.org/10.1093/bioinformatics/btu554 PMID: 25150250

31. Khan I. Open-source software for designing, documenting and reproducing biological experiments. 2014. 2015 Mar 19 [cited 4 Apr 2018]. http://protocolnavigator.org/index.html.

32. Research Data Alliance. 2018 Mar 26. [cited 2017 Aug. 16]. https://www.rd-alliance.org/about-rda

33. Executive Order—Making Open and Machine Readable the New Default for Government Information. (The White House Office of the Press Secretary, 2013).