Harvard Data Science Review • Issue 2.4, Fall 2020

Reproducibility in Science: A Metrology Perspective

Anne Plant[1] Robert Hanisch[2]

[1]Biosystems and Biomaterials Division, Materials Measurement Laboratory, National Institute of Standards and Technology, United States Department of Commerce, Gaithersburg, Maryland, United States of America,
[2]Office of Data and Informatics, Material Measurement Laboratory, National Institute of Standards and Technology, United States Department of Commerce, Gaithersburg, Maryland, United States of America

The MIT Press

Published on: Dec 16, 2020

DOI: https://doi.org/10.1162/99608f92.eb6ddee4

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

ABSTRACT

Scientific progress requires the ability of scientists to build on the results produced by those who preceded them. Because of this, there is concern that irreproducible scientific results are being reported. We suggest that while reproducibility can be an important hallmark of good science, it is not often the most important indicator. The discipline of metrology, or measurement science, describes a measurement result as a value and the uncertainty around that value. We propose a systematic process for considering the sources of uncertainty in a scientific study that can be applied to virtually all disciplines of scientific research. We suggest that a research study can be characterized by how sources of uncertainty in the study are reported and mitigated. Such activities can add to the value of scientific results and the ability to share data effectively.

Keywords: measurement science, reproducibility, sources of uncertainty, comparability, data, metadata

1. Introduction

Concern about what is commonly referred to as reproducibility of research results seems to be widespread across disciplines. Scientists, funding agencies and private and corporate donors, industrial researchers, and policymakers have decried a lack of reproducibility in many areas of scientific research, including computation (Peng, 2011), forensics (National Research Council, 2009), epidemiology (Ioannidis et al., 2005), and psychology (Open Science, 2015). Failure to reproduce published results has been reported by researchers in chemistry, biology, physics and engineering, medicine, and earth and environmental sciences (Baker, 2016). While there are sociological causes of the current concern, the principles and tools of metrology can provide some guidance for how to address this problem from a measurement science point of view.

Measurement science, or metrology, is the study of how measurements are made, and how data are compared. “Metrology is the science of measurement, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology” (BIPM). The purpose of this article is to highlight how measurement science is applied in the conduct of research in general, and in specific areas of research where reproducibility challenges have been noted. Measurement science has been traditionally applied to physical measurements. However, we suggest here that the thought process of measurement science is broadly applicable, even to purely theoretical studies, and will enable more effective data sharing. We will focus on tools and approaches for achieving measurement assurance, confidence in data and results, and the facility for sharing data.

1.1. Reproducibility, Uncertainty, and Confidence

Relevant definitions. The dictionary definition of the term *uncertainty* refers to the condition of being uncertain (unsure, doubtful, not possessing complete knowledge). It is a subjective condition because it pertains to the perception or understanding that one has about the value of some property of an object of interest. In measurement science, *measurement uncertainty* is defined as the doubt about the true value of a particular quantity subject to measurement (the ‘measurand’), and quantifying this uncertainty is fundamental to precision measurements (Possolo, 2015). The International Vocabulary of Metrology Metrology (VIM) (2012) (*International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*, 2012) is used by the international metrology community and provides definitions for many terms related to the current concerns about reproducibility. The definitions found in the VIM have been under development by an international community since the Metre Convention of 1875 which created the International Bureau of Weights and Measures (BIPM). While ‘reproducibility,’ ‘replicability,’ and related terms have been variously defined by different groups, the term ‘reproducibility’ has a precise definition in the international measurement science community. Table 1 lists a few of the terms in the VIM that describe the various aspects of a measurement process that relate to a discussion about confidence in scientific results.

Table 1. Some relevant terms and definitions that are consistent with the VIM. ‘Replicability’, a term that is often used in conjunction with the common use of ‘Reproducibility’, is not defined in the VIM.

Term	**Definition**	**Notes**
Reproducibility	Precision in measurements under conditions that may involve different locations, operators, measuring systems, and replicate measurements on the same or similar objects. The different measuring systems may use different measurement procedures.	A specification should give the conditions changed and unchanged, to the extent practical.
Repeatability	Precision in measurements under conditions that include the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time.	
Precision	Closeness of agreement between measured quantities obtained by replicate measurements on the same or similar objects under conditions of repeatability or reproducibility.	Usually expressed as standard deviation, variance or coefficient of variation.

Accuracy	Closeness of agreement between a measured quantity value and a true quantity value of a measurand.	

The role of reproducibility. In this article, we consider what reproducibility means from a measurement science point of view, and what the appropriate role of reproducibility is in assessing the quality of research. Measurement science considers reproducibility to be one of many factors that qualify research results. A systematic examination of uncertainties that can exist in the various components of research may provide a better alternative to assessing the reliability of research results than does a limited focus on reproducibility.

Reproducibility and the desire for confidence in research results. There are competing definitions associated with reproducibility that are commonly used (Pellizzari, Lohr, Blatecky, & Creel, 2017; Barba, 2018), and there have been a number of strategies taken to respond to the concern about reproducibility. Funding agencies, scientific journals, and private organizations have instituted checklists, requirements, and guidelines (Collins & Tabak, 2014; Nosek et al., 2015; R.F. Boisvert, 2016). There have been a number of sponsored activities focused on demonstrating the reproducibility of previously published studies by other laboratories ("Reproducibility Initiative," 2014; Weir, 2015). Checklists have met with some resistance (Baker, 2015), including the criticisms of the ‘one size fits all’ nature of the guidelines, that some of the criteria are inappropriate for exploratory studies, that the guidelines are burdensome to authors and reviewers, and that the emphasis on guidelines shifts the responsibility for scientific quality from scientists themselves to the journals. There are further concerns from funders and editors that they need to assume a policing role (Lash, 2015). Criticisms of the focus on reproducing results in independent labs cite the implicit assumption that only reproducible results are correct, and if a result is not reproducible it must be wrong. There are no easy answers for how to determine when the result of a complex study is sufficiently reproduced. Metrology laboratories spend significant effort in measurement comparisons, establishing consensus values, using reference materials, and determining confidence limits. This work is especially challenging when the measurements themselves are complicated or the measurand is poorly defined. From a practical point of view, the effort to reproduce published studies can be prohibitively expensive and time consuming (Maher, 2015).

The complexities associated with inter-laboratory reproducibility can be great, and when performed by metrology experts, inter-laboratory studies follow a formal and systematic approach (Maher, 2015). There is no doubt that demonstrating reproducibility of a result instills confidence in that result. But results can be reproduced and still be inaccurate (recall the many rapid confirmations of cold fusion, all of which turned out to be erroneous; see, for example, Mallove [1991]), suggesting that reproducibility is not a sufficient indicator of confidence in a result. In addition, a failure to reproduce is often the beginning of scientific discovery, and it may not be an indication that that any result is ‘right’ or ‘wrong.’ Particularly in the cases of complicated experiments, it is likely that different results are observed because different experiments are being conducted unintentionally. Without a clear understanding of what should be ‘reproducible,’ what variation in results is reasonable to expect, and what the potential sources of uncertainty are, it is easy to devote considerable resources to an unproductive goal.

An alternative to focusing on reproducibility as a measure of reliability is to examine a research result from the perspective of one’s confidence in the components of the study, and by acknowledging and addressing sources of uncertainty in a research study. Thompson (2017) goes further, suggesting that research methods should be reviewed and accredited as a prerequisite for submission for publication of research in journals. Uncertainty in measurement and transparency of research methods are unifying principles of measurement science and are critical for high quality research results.

1.2. The International Conventions of Metrology

The sources of variability in a measurement system and how they contribute to measurement uncertainty is an importance concept in measurement science. The National Institute of Standards and Technology (NIST), which is the national metrology institute (NMI) of the United States, and its 100-plus sister laboratories in other countries, promote these concepts for ensuring confidence in measurement results. The NMIs enable the intercomparability of measurement results worldwide, within the framework maintained by the International Bureau of Weights and Measures (*Bureau International des Poids et Mesures*, BIPM). These international efforts that underlie the intercomparability of measurement results in science, technology, and commerce and trade, have a long history, having enabled the development of modern physics beginning in the 19th century by the contribution of researchers including Gauss, Maxwell, and Thompson (BIPM). The work in metrology at national laboratories impacts international trade and regulations that assure safety and quality of products, advances technologies to stimulate innovation and to facilitate the translation of discoveries into efficiently manufactured products, and in general serves to improve the quality of life. The concepts and technical devices that are used to characterize measurement uncertainty evolve continuously to address emerging challenges as an expanding array of disciplines and sub-disciplines in chemistry, physics, materials science, and biology are considered.

While the concepts of metrology are a primary responsibility of national measurement laboratories, the goal is that these concepts should be widely applicable to all kinds of measurements and all types of input data. As an example of their potential universality, the terms of the VIM have been explicitly adapted to provide a useful guide for geoscience research (Potts, 2008).

2. Indicators of Confidence and Reduction of Uncertainty in Research Results

Sources and quantification of uncertainty. Reproducibility is one of the concepts considered when the metrology community assesses measurement uncertainty, but it is not the only one. Uncertainties in measurement typically arise from multiple sources. In the Guide to Uncertainty in Measurement (BIPM, 2008), the international metrology community lists a number of examples of sources of uncertainty (see Table 2).

Table 2. Possible sources of uncertainty in a measurement (from the Guide to the Expression of Uncertainty in Measurement (GUM), Section 3.3.2 (BIPM, 2008). These sources are not necessarily independent, and some of sources 1 to 9 may contribute to source 10. Of course, an unrecognized systematic effect cannot be taken into account in the evaluation of the uncertainty of the result of a measurement but nevertheless contributes to its error.

1. Incomplete definition of the measurand;
2. Imperfect realization of the definition of the measurand;
3. Non-representative sampling—the sample measured may not represent the defined measurand;
4. Inadequate knowledge of the effects of environmental conditions on the measurement or imperfect measurement of environmental conditions;
5. Personal bias in reading analogue instruments;
6. Finite instrument resolution or discrimination threshold;
7. Inexact values of measurement standards and reference materials;
8. Inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm;
9. Approximations and assumptions incorporated in the measurement method and procedure;
10. Variations in repeated observations of the measurand under apparently identical conditions.

The sources of measurement uncertainty can be systematically identified and quantified. For a discrete measurement, such as quantifying the amount of a substance, statistical measures of uncertainty in the measurement are compared across metrology laboratories to assess their relative confidence in the measurement. Uncertainties are determined in each laboratory at each step of the measurement process and might include, for example, the error in replicate weighing and pipetting steps. An expanded uncertainty budget is determined as an aggregate value that accounts for the combination of uncertainties at all steps in a measurement process. The quantification of uncertainty provides a basis for the limits within which that measurement, or deviation from that measurement, is meaningful.

In measurement science, a measurement consists of a value and the uncertainty around that value. The calculation of an expanded uncertainty takes into account all sources of uncertainty at every stage of the measurement. In a research setting, the formalism of such a calculation is rarely necessary, but acknowledging and addressing sources of uncertainty are critical. Regardless of discipline, at each step of a scientific study we should be able to identify the potential sources of uncertainty, including measurement uncertainty, and report the activities that went into reducing the uncertainties inherent in the study. One might argue that the testing of assumptions and the characterization of the components of a study are as important to report as are the ultimate results of the study.

Systematic reporting of sources of uncertainty. While research reports typically include information about reagents, control experiments, and software, this reporting is rarely as thorough as it could be, and the presentation of such details is not systematic. We have suggested a systematic framework (Plant et al., 2018) (shown in Table 3) for identifying and mitigating uncertainties that includes explanation of assumptions made, characteristics of materials, processes, and instrumentation used, benchmarks and reference materials, tests to evaluate software, alternative conclusions, etc. These data and metadata are critical to reducing the ambiguity of the results. Table 3 is a general guide that is applicable to most areas of research.

Table 3. Identifying, reporting, and mitigating sources of uncertainty in a research study (Plant et al., 2018).

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 (Plant, Becker, Hanisch, et al., 2018) and other sources (Rosslein, Elliott, Salit, et al., 2015).

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.

If we assume that 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. The supporting evidence that defines the characteristics of the data and analysis, and tests the assumptions made, provides additional confidence that one has in the results. Confidence is established when supporting evidence is provided about assumptions, samples, methods, computer codes and software, reagents, analysis methods, etc., that went into generating a scientific result. Confidence in these components of a study can be an indication of the confidence we can have in the result. Confidence can be increased by recognizing and mitigating sources of uncertainty.

3. Metrology Tools for Achieving Confidence in Research Results

The systematic consideration of sources of uncertainty in a research study such as presented in Table 3 can be aided by a number of visual and experimental tools. For example, an *experimental protocol* can be graphed as a series of steps, allowing each step to be examined for sources of uncertainty. This kind of assessment can be valuable for identifying activities that can be optimized, or places where *in-process controls* or *benchmarks* can be used to allow the results of intermediate steps and performance of the instrument to be evaluated before proceeding. Another useful tool is an Ishikawa or cause-and-effect diagram (Rouse, 2015). This is a systematic way of charting all the experimental variable that might contribute to uncertainty in the result.

Below are some of the services and products that NIST supplies that help practitioners realize some of the concepts that are itemized in Table 3.

Reference materials. Instrument performance characterization and experimental protocol evaluation are aided by the use of Reference Materials (RMs) and Standard Reference Materials® (SRMs). SRMs are the most highly characterized reference materials produced by NIST; RMs can be produced more quickly and are fit for purpose. RMs and SRMs are developed to enhance confidence in measurement by virtue of their well-characterized composition or properties, or both. RMs are supplied with a certificate of the value of the specified property, its associated uncertainty, and a statement of metrological traceability. These materials are used to determine instrument performance characteristics, perform instrument calibrations, verify the accuracy of specific measurements and support the development of new measurement methods by providing a known sample against which a measurement can be compared. Instrument design and environmental conditions can be systematic sources of uncertainty that the use of reference materials with highly qualified compositional and quantitative characteristics can help identify. Reference materials also assist the evaluation of experimental protocols and provide a known substance that can allow comparison of results between laboratories. NIST SRMs are often used by third-party vendors who produce reference materials to provide traceability to a NIST certified value. A NIST Traceable Reference Material^{TM} has a well-defined traceability link to existing NIST standards (May et al., 2000). Examples of reference materials produced by NIST include serum containing precise amounts of metabolites and hormones, reference genomes for determining accuracy of DNA sequencing, silicon implanted with boron for depth profiling, and materials with certified thermoelectric properties.

Any lab can create their own reference materials that are appropriate for their specific purpose. These materials should be homogeneous (i.e. can be sampled representatively) and be stable over the time frame in which they are to be used.

Calibration services. NIST (2019) provides the highest order of calibration services for instruments and devices available in the United States satisfying the most demanding and explicit criteria for quality assurance.

These measurements directly link a customer's precision equipment or transfer standards to national and international measurement standards.

Reference instruments. NIST supports accurate and comparable measurements by producing and providing Standard Reference Instruments. Reference instruments allow customers the ability to make reference measurements or generate reference responses in their facilities based on specific NIST reference instrument designs. These instruments support assurance of measurements of time, voltage, temperature, etc.

Underpinning measurements that establish confidence. RMs and SRMs, Calibration Services, and Standard Reference Instruments provide confidence in primary measurements, and also in the instruments and materials that underpin the primary laboratory or field measurement, such as temperature sensors, pH meters, photodetectors, and light sources.

Interlaboratory comparison studies. NIST leads and participates in Interlaboratory comparison studies as part of their official role in the international metrology community (BIPM), and in less formal studies. An example of a less formal study involving NIST was a comparison with five laboratories to identify and mitigate sources of uncertainty in a multistep protocol to measure the toxicity (EC_{50}) of nanoparticles in a cell-based assay. The study was undertaken because of the large differences in assay results and conclusions from the different labs, and the inability of the participants to easily identify and control the sources of uncertainty that resulted in the observed irreproducibility. A *cause-and-effect diagram* was created to identify all potential sources of uncertainty, and this was followed by a preliminary study of that used a *design of experiment* approach to perform a sensitivity analysis to determine how nominal variations in assay steps influenced the EC_{50} values (Rosslein et al., 2015). Two variables were identified as particularly important to specify, and these were systematically explored for their effect. As a result of the analysis, a *series of in-process controls* were run with every measurement. The results of the control wells were expected to be within a specified range to assure confidence in the test result. Control wells assess variability in pipetting, cell retention to the plate after washing, nanoparticle dispersion, and other identified sources of variability. The outcome was a robust protocol, benchmark values for intermediate results, concordant responses in EC_{50} to a reference preparation by all laboratories, and confidence in the meaningfulness of the results reported in each laboratory.

In general, laboratories that participate in formal inter-laboratory studies (Hibbert, 2007) know from experience that it often takes several iterations of studies, and intensive determination of sources of variability, before different expert laboratories produce comparable results. The result of these efforts is a more robust and reliable experimental protocol in which critical parameters are controlled.

Standard reference data. The NIST Standard Reference Data portfolio comprises nearly 100 databases, tables, image and spectral data collections, and computational tools that have been held to the highest possible level of *critical evaluation*. Many of these are compilations of data published in journals that are reviewed and assessed for measurement practices and uncertainty characterization by NIST or NIST-contracted topic experts. Others consist of measurements made by NIST scientists and validated through inter-laboratory comparisons.

Specifically, *critical evaluation* means that the data are assessed by experts and are trustworthy such that people can use the data with confidence and base significant decisions on the data. For numerical data, the critical evaluation criteria are:

> a. Assuring the *integrity* of the data, such as provision of uncertainty determinations and use of standards;
>
> b. Checking the *reasonableness* of the data, such as consistency with physical principles and comparison with data obtained by independent methods; and
>
> c. Assessing the *usability* of the data, such as inclusion of metadata and well-documented measurement procedures.

For digital data objects, the critical evaluation criteria are:

> a. Assuring the object is *based on* physical principles, fundamental science, and/or widely accepted standard operating procedures for data collection; and
>
> b. Checking for *evidence* that
>   i. The object has been *tested*, and/or
>   ii. Calculated and experimental data have been *quantitatively compared*.

NIST SRD criteria serve as an exemplar of the kind of processes that, if adopted more widely, would improve confidence in research data generally.

4. Metrological Caveats to Reproducibility

Definitional challenges associated with reproducibility. When national metrology laboratories around the world compare their measurement results in the formal setting of the BIPM, there are accepted expectations regarding expression of uncertainties in the measurements reported, and how the measurements from different laboratories are compared. The reporting of the values and uncertainties from the different labs provides an indication of relative proficiency that can be accessed for comparative purposes. Outside of this formal setting, it is less clear how exactly to compare results from different laboratories, and therefore, how to assess whether a result was reproducible or not. Many of our greatest measurement challenges today preclude an easy assessment of reproducibility. The examples below are the kinds of dilemmas that inspired the NASEM Report on Reproducibility and Replicability in Science to define ‘replicability’[1] as distinct from ‘reproducibility’ (National Academies of Sciences & Medicine, 2019). Below, we suggest how concepts associated with evaluation of uncertainty might assist assessment of concordance in research results that are difficult to compare.

Identity vs. a numerical value. While DNA sequencing is not the only case, it is a good example of where the identity of the bases and their relative locations *is the measurand*. A NIST-hosted consortium called Genome in a Bottle (GIAB) (The Joint Initiative for for Metrology in Biology, 2016) has been working for several years to amass sufficient data that would allow an evaluation of the quality of data that can be achieved by different laboratories. This is a large inter-laboratory effort in which the same human DNA material is analyzed with different instruments and using different bioinformatics pipelines. The data indicate that good concordance of sequence is achieved readily in some portions of the genome, and other regions are more problematic and require accumulation of more data. In other regions, where there is a large number of repeated sequences for example, it may be impossible to establish a high level of confidence. Putting a numerical value on concordance under these circumstances is challenging. The comparison of data across laboratories is critical to establish the characteristics of the sequences for which good concordance is possible.

Complexity of research studies and measurement systems. Part of the challenge in genome sequencing, which is under investigation in GIAB, is that instruments used to sequence DNA have different biases, different protocols introduce different biases, and the software routines for assembling the intact sequence from the fragments often give different results. Determining the sources of variability and whether it is even possible to calculate an uncertainty is still ongoing. For many measurements associated with complex research studies, making a detailed uncertainty determination is in itself a research project. However, reporting what is known about each of the sources of uncertainty presented in Table 3 would be possible, and should be encouraged.

No ground truth. GIAB is a good example that has much in common with many of our most pressing measurement challenges today. Even with a reference material that everyone can use and compare the results from, the real answer is unknown or unknowable; i.e., there is no ground truth sequence. DNA sequencing is certainly not the only example of this dilemma. The best that one can do is determine a consensus answer, i.e., a value that most of the community would come to (or close to). As with the other examples, interlaboratory comparison of data and complete reporting of sources of uncertainty provide confidence in the results.

How close is close enough to call reproducible? Establishing that a result has been reproduced or not can be complicated. Especially when different instrumentation is used, the exact value of a complex measurement may not be identical to that achieved by another laboratory. If an expanded uncertainty was determined, as is done when national metrology laboratories compare their measurements, then a comparison could be made, but this may not happen in a typical research environment given the complicated nature of many of the studies being performed. Human cell line authentication is an example where a committee had to arbitrarily establish a threshold of similarity in the identification of the size and number of short tandem repeat (STR) sequences. Above 75% concordance in STR sequences identified was determined to be sufficient for identification (American Type Culture Collection Standards Development Organization Workgroup, 2010). This threshold reflects the incomplete state of understanding of the sources of differences in the sequence lengths, not necessarily to poor quality of the raw data.

Unique events, sparsity of data. Numerous scientific inquiries rely on observations of one-time events: earthquakes, tsunamis, hurricanes, epidemics, supernovae, etc. Researchers gain understanding of such phenomena through observations of multiple distinct events having similar, but not identical, behavior. In these kinds of studies, it is most critical to evaluate and report sources of uncertainty in the measurements if the measurements are to be compared to one another.

5. Metadata Issues

Enabling reuse of results by establishing confidence in assumptions, software, and data. It is hard to imagine that any experimental research result in the present era that does not rely on computer software, ranging from spreadsheets to shared community software packages to complex custom codes. Too often research papers include the throw-away line ‘the data were reduced in the usual manner,’ and no record of the various input parameters and options is provided. As noted by Stodden and Miguez (2014), documenting what software was used and sharing code are essential practices for assuring reproducible and reusable research. Fortunately newer practices are facilitating this through the publication and sharing of data and processing steps in, for example, Jupyter notebooks, and the registration of software packages and source codes in shared indexes (e.g., the Astrophysics Source Code Library, 2018) or the NIST Materials Resource Registry (Becker et al., 2017). In fact, the Materials Resource Registry indexes both data and software, treating the latter as a special type of data.

Most researchers would agree that researchers should share data and software, including source code. The adoption of the FAIR principles of findable, accessible, interoperable and reusable (Wilkinson et al., 2016) is helping to make the effective sharing of data possible. Ultimately, however, the ability to build on published research results will be limited by the reliability of the data, assumptions, and software on which the conclusions are based. It should be *de rigueur* to demonstrate confidence in these components of a study by providing supporting evidence. Outside of computer science, the unreliability of software is often underappreciated, although there are efforts to make the biological imaging community more aware that image analysis algorithms are not all equivalent and do not perform equally well on all images (Dima et al., 2011; Bajcsy et al., 2015; Caicedo et al., 2017). Rigorous testing of software should be performed, as it has long been understood that numerical software has reliability challenges (Boisvert et al., 1997). To ensure our results are generalizable beyond a particular computing environment, we should test if the results can be reproduced from computer code running on different machines under different operating systems but with the same inputs (Mytkowicz, Diwan, Hauswirth, & Sweeney, 2009; Blackburn, 2016).

Sharing of data and software within and across disciplines should be a strong motivator for adopting a framework of general principles for assessing confidence in research studies. If researchers, for example, are going to use laboratory data as input, the details of the experiment and the extent to which the data were qualified might influence model selection and details associated with the study, including the effect of propagating measurement uncertainty. Particularly when considering the use of data in interdisciplinary research, it is important that the quality of the data generated in one field is understood by a user of that data who may be not be an expert in that field of study. Identifying criteria that establish confidence in results that everyone understands will facilitate appropriate reuse of study results.

Availability of data, metadata, and provenance information. As our ability to store, transfer, and mine large amounts of data improves, the importance of establishing confidence in the quality of those data increases. At the moment, there are few tools for assessing quality of data. One project underway is focused on identifying the presence of supporting data out of published research reports (McIntosh, 2017). In addition, NIST's Thermodynamics Data Center has long employed partially automated data quality assessment tools (Frenkel et al., 2005). Adoption of a widely accepted systematic framework for reporting such data would enable this effort. Supporting data that provides confidence in assumptions, models, experimental data, software and analysis needs to be collected more diligently and reported more systematically. Particularly difficult is the collection and reporting of details of protocols used in studies that involve complex experimental systems. Improved metadata acquisition software incorporated into laboratory information management systems could facilitate the collecting, sharing, and reporting of details of protocols. The Research Data Alliance has recently started a new Working Group on Persistent Identification of Instruments (2017), which for experimental data could greatly improve provenance through tracing data back to a particular instrument and its associated calibration information. Expert software systems that facilitate the collection of highly granular experimental metadata could help to identify subtle experimental differences that are sources of uncertainty and causes of irreproducibility; this knowledge might provide important information about the systems under study. A requirement for effective metadata sharing is the development of better methods of harmonized vocabularies possibly through the use of natural language methods (Bhat, Bartolo, Kattner, Campbell, & Elliott, 2015). Unambiguous meanings and context in metadata labels would enable searching and discovery of similar and dissimilar experimental protocol details (Gregory, Groth, Scharnhorst, & Wyatt, 2020). Within the metrology community there is the concept of 'fit for purpose.' Good metadata will make it clear whether a dataset is relevant and appropriate for use, e.g., noting its extent of applicability, reliability, and uncertainty.

How much reporting is enough? Failure to reproduce a result can play a critical role in discovery of imperfect measurements or observations and can uncover fundamental flaws in theoretical assumptions and interpretations. An example is the use of the Hubble Constant to determine the age of the universe, which in the 1930s was inconsistent with the determination from radioactive dating on Earth (which indicated the age of the Earth exceeded the age of the Universe!). Twenty years later it was found that the calibration of the distance scale (based on the period-luminosity relationship for Cepheid variable stars) was applied mistakenly to star clusters rather than individual stars (Huchra, 2008). Often, failure to reproduce is the result of failure to identify and control major sources of variability. This can indicate that some parameter that has not been controlled is an important source of uncertainty (Thompson & Ellison, 2011). In biomedical research, there can be so many uncontrolled and hidden variables that there is a high likelihood that experiments performed in different labs are actually substantially different. If there was full and systematic reporting of experimental details, it may be possible to discover previously unrecognized sources of variability that provide important scientific insight. One could argue that it is impossible to eliminate bias and to report every experimental variable, protocol nuance, instrument parameter, etc. One could also argue that doing better than is currently done would increase the rate at which scientific advances occur. More investment in software tools to enable the collection, storage, and searching of metadata would improve our abilities to more fully describe our research studies.

6. Discipline-Specific Considerations

The importance of reproducibility and data sharing relative to other aspects of the scientific process can be different for different scientific disciplines. Below are some brief examples intended to highlight similarities and differences associated with data reproducibility and sharing for different fields of study. Regardless of discipline, at each step of a scientific endeavor, we should be able to ascertain the activities that went into testing assumptions and characterizing components of the study.

6.1. Astronomy

With the advent of large-scale digital sky surveys and routine pipeline-based calibrations that produce science-ready data products, the vast majority of astronomical research data is open (often after a nominal proprietary period such as 12 months). These advances have led to substantial reanalysis and repurposing of data (nearly two-thirds of peer reviewed publications based on Hubble Space Telescope observations are based on archival data) ("Mikulski Archive for Space Telescopes"). The Sloan Digital Sky Survey (SDSS) has yielded some 8,000 peer-reviewed publications, the vast majority of which have been written by researchers who are not part of the SDSS project ("Astrophysics Data System"). Reproducibility problems in astronomy are relatively rare, owing to the prevalence of open data in astronomy, the wide use of standard software packages and pipeline-calibrated data, and a relatively small and well-connected research community (~10,000 professional astronomers worldwide). Where they do exist, as in the Hubble constant studies mentioned earlier, they often result from incomplete information about the phenomenon being measured.

6.2. Physics

Despite the strong theoretical footing associated with research in physics, this scientific discipline is not free from reliability issues. For large-scale high-energy physics experiments such as those at the Large Hadron Collider, one expects that extreme care has been taken in acquiring, calibrating, and analyzing the data. But even big experiments can produce erroneous results, such as was the case in 2011 when the OPERA experiment in Italy reported the preliminary finding that neutrinos produced at CERN travelled faster than the speed of light (Brumfiel, 2012). We would expect that there are many small laboratory experiments in physics that have problems similar to those in other disciplines, e.g., where instrumental metadata is stored in proprietary vendor formats that are not easily interpreted and where hidden variables lead to challenges in reproducibility. However, the level of theoretical understanding that has developed over centuries makes physics less susceptible than other fields to reproducibility failures.

6.3. Materials Science

Measurements in materials science have recently received additional attention through the national Materials Genome Initiative (MGI), whose goal is to accelerate the development of new materials at lower cost through better integration of computer simulation and experimentation. There are significant challenges in reproducibility in materials science largely around growth, processing, and sample preparation and processing. For example, the fine-scale structure in an alloy can vary greatly depending on how it is cooled. Complexity in materials systems such as nanocomposites, where homogeneity might be lacking and where interfacial properties are poorly understood, proposes a grand challenge in terms of experimental reproducibility that compromises the lab-to-market pathway. The disruptive promise from novel materials systems such as graphene and 2D systems is often discouraged by poor reproducibility from growth and processing, which underlies limited understanding of the physico-chemical phenomena underpinning those events. Reproducibility, and the development of predictive models, suffer when the growth and processing history of a material is not fully documented, when unknown (and hence unmeasured) effects impact properties, or when significant instrumental parameters are hidden in proprietary binary data formats.

6.4. Biology

For highly complicated studies that involve a very large number of parameters such as those conducted in the biomedical sciences, it may be very difficult to uncover, and impossible to control, all sources of uncertainty and variability in a study. In such cases an inability to reproduce a result may simply indicate that the two experiments were in fact different, possibly for reasons that are not well understood. For these kinds of studies, it would be of great importance to have sufficient information and facility to compare exactly which aspects of which steps in the processes are different; such a meta-analysis may provide valuable scientific insight. In addition to the challenges of parameter space, many biological and biomedical systems are characterized by a degree of complexity that is not apparent in other sciences. For example, the importance of stochastic fluctuations in biochemical reactions within cells, the number of biochemical processes that can be involved in cellular response, and the promiscuity of alternative intracellular pathways by which environmental information can be processed, can result in highly complex and heterogeneous biological responses that make reproducibility very difficult to define. Biological heterogeneity is different from, but is convoluted with, measurement noise. Accurate evaluation of biological heterogeneity requires independent assessment of measurement uncertainty. The reporting of statistical means for biological data is common but may not very informative because of this convolution. Also, the mean is often not an adequate metric since biological response functions are rarely well-described by simple symmetric distributions. Table 4 articulates some challenges and strategies in single cell experiments (Keating et al., 2018) for distinguishing measurement uncertainty from biological variability.

While techniques like design of experiment can be used to assess interactions between multiple variables that are sources of variability in measurement, we are just now entering an era where the complexity of the biological systems under study, not just the experiments, can be addressed. In the realm of cell biology for example, complex control mechanisms involve many molecular species and have both temporal and spatial dependencies. Our ability to collect, store, search and share very large data sets and their provenance will be instrumental to recognizing the patterns of events in complex systems and for developing the understanding of fundamental principles for predicting their outcomes. More than ever, we must have confidence in the data that will be available for development of models of such complex systems.

Table 4. Distinguishing measurement uncertainty from biological variability in a single cell assay (Keating et al., 2018).

Challenge		Strategy
Measurements of biological response to environmental conditions		• Measure sufficient numbers of cells to assure adequate sampling of population diversity (heterogeneity) • Use appropriate statistics for comparison (e.g., cumulative distributions, not means) • Both the mean response and teh shape of the distribution of responses may change in response to treatment. • Use appropriate positive and negative controls. • Compare teh results from orthogonal analytical methods: different methods should return similar responses. • Measure response function (concentration or time dependence) to test for a systematic effect.

Distinguish inherent biological heterogeneity from measurement variability	- Measurement variability	- Quantify the uncertainty due to variability (e.g., SD) in the measured value due to instrument response. Measure within day (repeatability) and day-to-day (reproducibility). - Test the sources of measurement variability (technicians, reagents, environment, algorithms, protocols), and try to mitigate them. - Quantify the variation in results from the same sample on different platforms.
	- Biological heterogeneity due to stochasitc fluctuations	- Test the stability of the ddistribution of the population characteristic or phenotype. - Measure similar distributions from repeated measurements of the population over long time intervals. - Sorted “subpopulations” will relax over time in culture to a stable distribution similar to the original distribution. - “Subpopulations” are genetically identical.
	- Biological heterogeneity due to genetic/genomic differences	- Population phenotypic heterogeneity diverges over time in culture. - Subpopulations have transcriptomic and genomic differences.

Minimize uncertainty in measurement variability		• Assess instrument performance with benchmarking materials for signal to noise, linearity of response, limit of detection, and saturation. • Use control materials (e.g., spike-in RNA into transcriptomic samples) to test and compare assay platform response and to assess technical proficiency. • Use control materials to test and optimize protocols for accuracy, precision, sufficient dynamic range, sensitivity, specificity, and robustness to small protocol changes. • Test and compare algorthims for robustness and accuracy against ground truth (if available).

7. Qualifying and Characterizing Measurement Systems

Table 5 is based on criteria provided by the U.S. Food and Drug Administration for qualifying assays that are used to characterize a regulated biological product. The measurement elements in the table are criteria that, when identified, help to provide confidence about the measurement system and the results (Plant, Locascio, May, & Gallagher, 2014). While these best practices are directed at measurements of biological systems, they are sufficiently general to be applicable to most experimental situations. Achieving the knowledge of these measurement elements requires a high level of understanding of, and experience with, the measurement system. Reporting these characteristics for a measurement system has advantages for both the experimentalist and for the user of the resulting data by providing documented evidence that there is a high level of confidence in the accuracy of the resulting measurements. It is critical that if an assay indicates an unexpected result, it is a result that is an accurate characterization of the material itself; one doesn’t want to question if the assay itself might be flawed. Reporting the qualifying characteristics of the measurement method helps to establish confidence in research results.

Table 5. Key elements of a good measurement. Adapted from Plant et al. (2014).

Measurement element	Description	Best practice

Accuracy	The measurement delivers the true value of the intended analyte (i.e., the measurand)	Test your experimental observation using orthogonal analytical methods. Use well-defined reference materials to check instrument response and method validity.
Precision	Repeatability (replicates in series) and reproducibility on different days and in different labs	Replicate the measurement in your own lab, perhaps with different personnel. Have another lab perform the experiment. Participate in an inter-laboratory comparison study.
Robustness	Lack of sensitivity to unintended changes in experimental reagents and protocols	Test different sources of reagents, fixation conditions, incubation times, cell densities and analysis software
Limit of detection	Given the noise in the measurement, the level below which the response is not meaningful	Use appropriate positive and negative controls to determine background signal, and use dispersion in replicate measurements to determine measurement uncertainty.
Response function	Dependence of signal on systematic change in experimental condition	Systematically test concentration or activity with reference samples; determine the range in which the assay is sensitive.
Specificity	The analytical result is not confounded by sample composition or physical characteristics	When testing samples from different sources, ensure that apparent response differences are not due to sample matrix differences by using spike-in controls.

8. Summary Considerations for Data Science

A workshop entitled “Improving Reproducibility in Research: The Role of Measurement Science” was hosted by the National Physical Laboratory (NPL), Teddington, UK, in May of 2018 and was co-organized by NPL, NIST, and several other NMIs (Hanisch, Gilmore, & Plant, 2019). Some of the most important conclusions from the workshop regarding steps that could be taken to enable confident datasharing include:

- Develop and deploy tools that make it easier to collect and document experimental protocols (laboratory information management systems, metadata extractors, Jupyter notebooks).
- End practices such as p-hacking (Bishop, 2019), a posteriori data filtering, etc., through improved education in statistics and data handling.

- Verify/qualify the software used in support of experiments and analysis.
- Promote data stewardship and software development activities as career positions integral to the advancement of science.
- Establish long-term institutional commitments to data preservation and dissemination.
- Apply the FAIR (making data Findable, Accessible, Interoperable, and Reusable) principles to research data broadly (Wilkinson et al., 2016).
- Develop and gain community adoption of discipline-based metadata standards, with mappings to complementary research domains.
- Develop techniques for quantifying the uncertainties and understanding the results of machine learning and deep learning algorithms; provide domain-specific ground-truth datasets.
- Engage with publishers and editors of scholarly journals to work toward better presentation of full provenance of research, including the development of machine-actionable research reports and the reporting of negative results.

9. Conclusions

In an era where there are many print and electronic journals for publishing scientific results, and facility for storing and sharing large amounts of data electronically, we have an unprecedented opportunity to advance our collective knowledge of the natural world. Explicitly applying concepts associated with the science of metrology to the practice of scientific research more broadly could have a profound effect on the quality of research by increasing confidence in data and enabling effective data sharing.

We suggest that in research planning, proposal evaluation, and review of research reports, science may be better served if we place a greater emphasis on identifying the sources of uncertainty in the studies than on the reproducibility of the results. However, we should also emphasize that irreproducibility of research results is not necessarily indicative of bad science, and that disagreement between laboratories often arises because not all aspects affecting the measurement are known. Arguably, it is through such inconsistencies that science advances.

Disclosure Statement

An earlier version of this work has appeared online at https://www.nap.edu/resource/25303/Metrology%20Perspective%20on%20Reproducibility.pdf

References

American Type Culture Collection Standards Development Organization Workgroup, A. S. N. (2010). Cell line misidentification: The beginning of the end. *Nature Reviews Cancer*, *10*(6), 441–448. https://doi.org/10.1038/nrc2852

Astrophysics Data System. Retrieved from https://ui.adsabs.harvard.edu/

Astrophysics Source Code Library. (2018). Retrieved from ASCL.net

Bajcsy, P., Cardone, A., Chalfoun, J., Halter, M., Juba, D., Kociolek, M., … Brady, M. (2015). Survey statistics of automated segmentations applied to optical imaging of mammalian cells. *BMC Bioinformatics, 16*, 330. https://doi.org/10.1186/s12859-015-0762-2

Baker, M. (2015). US societies push back against NIH reproducibility guidelines. *Nature*. https://doi.org/10.1038/nature.2015.17354

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. https://doi.org/10.1038/533452a

Barba, L. A. (2018). Terminologies for reproducible research. *arXiv.* https://doi.org/10.48550/arXiv.1802.03311

Becker, C. A., Dima, A., Plante, R. L., Youssef, S., Medina-Smith, A., Bartolo, L. M., … Brady, M. C. (2017). Development of the NIST Materials Resource Registry as a means to advertise, find, and use materials-related resources. *Materials Science and Technology Society*.

Bhat, T. N., Bartolo, L. M., Kattner, U. R., Campbell, C. E., & Elliott, J. T. (2015). Strategy for extensible, evolving terminology for the Materials Genome Initiative efforts. *JOM, 67*(8), 1866–1875. https://doi.org/10.1007/s11837-015-1487-4

BIPM. (n.d.). *Brief history of the SI*. Retrieved from http://www.bipm.org/en/measurement-units/history-si/

BIPM. (n.d.). *Metrology*. Retrieved from https://www.bipm.org/en/worldwide-metrology/

BIPM. (2008). Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Retrieved from https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature, 568*(7753), 435. https://doi.org/10.1038/d41586-019-01307-2

Blackburn, S. M. D., Hauswirth, M., Sweeney, P. F., Amaral, J. N., Brecht, T., Bulej, L., Click, C., Eeckhout, L., Fischmeister, S., Frampton, D., Hendren, L. J., Hind, M., Hosking, A. L., Johnes, R. E., Kalibera, T., Keynes, N., Nystrom, N., & Zeller, A. (2016). The truth, the whole truth, and nothing but the truth: A pragmatic guide to assessing empirical evaluations. *ACM Transactions on Programming Languages and Systems, 38*(4). Article 15. https://doi.org/10.1145/2983574

Boisvert, R. F. (2016). Incentivizing reproducibility. *Communications of the ACM, 59*(10), 5. https://doi.org/10.1145/2994031

Boisvert, R. F., & International Federation for Information Processing. (1997). *Quality of numerical software: Assessment and enhancement* (1st ed.). London, New York: Published by Chapman & Hall on behalf of the International Federation for Information Processing.

Brumfiel, G. (2012). Neutrinos not faster than light. *Nature*. https://doi.org/10.1038/nature.2012.10249

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., … Carpenter, A. E. (2017). Data-analysis strategies for image-based cell profiling. *Nature Methods, 14*(9), 849–863. https://doi.org/10.1038/nmeth.4397

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature, 505*(7485), 612–613. https://doi.org/10.1038/505612a

Dima, A. A., Elliott, J. T., Filliben, J. J., Halter, M., Peskin, A., Bernal, J., … Plant, A. L. (2011). Comparison of segmentation algorithms for fluorescence microscopy images of cells. *Cytometry A, 79*(7), 545–559. https://doi.org/10.1002/cyto.a.21079

Frenkel, M., Chirico, R. D., Diky, V., Yan, X. J., Dong, Q., & Muzny, C. (2005). ThermoData engine (TDE): Software implementation of the dynamic data evaluation concept. *Journal of Chemical Information and Modeling, 45*(4), 816-838. https://doi.org/10.1021/ci0500067b

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review, 2*(2). https://doi.org/10.1162/99608f92.e38165eb

Hanisch, R. J., Gilmore, I. S., & Plant, A. L. (2019). Improving reproducibility in research: The role of measurement science. *Journal of Research of the National Institute of Standards and Technology, 124*, Article 124024. https://doi.org/10.6028/jres.124.024

Hibbert, D. B. (2007). *Quality assurance for the analytical chemistry laboratory.* Oxford University Press

Huchra, J. P. (2008). THE HUBBLE CONSTANT. *2016*. Retrieved from https://www.cfa.harvard.edu/~dfabricant/huchra/hubble/

International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (JCGM 200:2012). (2012). Retrieved from http://www.bipm.org/en/publications/guides/vim.html

Ioannidis, J. P., Bernstein, J., Boffetta, P., Danesh, J., Dolan, S., Hartge, P., … Khoury, M. J. (2005). A network of investigator networks in human genome epidemiology. *The American Journal of Epidemiology, 162*(4), 302–304. https://doi.org/10.1093/aje/kwi201

The Joint Initiative for for Metrology in Biology. (2016). *Genome in a bottle (GIAB)*. Retrieved from http://jimb.stanford.edu/giab/

Jupyter. (2018). Retrieved from http://jupyter.org/

Keating, S. M., Taylor, D. L., Plant, A. L., Litwack, E. D., Kuhn, P., Greenspan, E. J., … Kuida, K. (2018). Opportunities and challenges in implementation of multiparameter single cell analysis platforms for clinical translation. *Clinical and Translational Science*, *11*(3), 267–276. https://doi.org/10.1111/cts.12536

Lash, T. L. (2015). Declining the transparency and openness promotion guidelines. *Epidemiology, 26*(6), 779–780. https://doi.org/10.1097/EDE.0000000000000382

Maher, B. (2015). Cancer reproducibility project scales back ambitions. *Nature*. https://doi.org/10.1038/nature.2015.18938

Mallove, E. F. (1991). *Fire from ice: Searching for the truth behind the cold fusion furor*. New York, N.Y.: J. Wiley.

Materials Genome Initiative. (n.d.). Retrieved from https://www.mgi.gov/

May, W. E., Parris, R., Beck, C., Fassett, J., Greenberg, R., Guenther, F., … MacDonald, B. (2000). *Definitions of Terms and Modes Used at NIST for Value-Assignment of Reference Materials for Chemical Measurements*. Gaithersburg, MD Retrieved from https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication260-136.pdf

McIntosh, L., Juehne, A., C. H., Vitale, Liu, X., Alcoser, R., Lukas, J. C., & Evanoff, B. (2017). Repeat: A framework to assess empirical reproducibility in biomedical research. *OSFPreprints*. https://doi.org/10.17605/OSF.IO/4NP66

Mikulski Archive for Space Telescopes. (n.d.). Retrieved from https://archive.stsci.edu/hst/bibliography/pubstat.html

Mytkowicz, T., Diwan, A., Hauswirth, M., & Sweeney, P. F. (2009). Producing wrong data without doing anything obviously wrong! *SCM Sigplan Notices*, *44*(3), 265–276. https://doi.org/10.1145/1508244.1508275

National Academies of Sciences, Engineering, & Medicine. (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press.

NIST Calibration Program Calibration Services Users Guide SP 250 Appendix. (2019). Gaithersburg MD. Retrieved from https://www.nist.gov/sites/default/files/documents/2019/02/11/feeschedule2019.pdf

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pellizzari, E. D., Lohr, K. N., Blatecky, A. R., & Creel, D. R. (2017). *Reproducibility: A primer on semantics and implications for research*. Research Triangle Park, NC: RTI Press.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Plant, A. L., Becker, C. A., Hanisch, R. J., Boisvert, R. F., Possolo, A. M., & Elliott, J. T. (2018). How measurement science can improve confidence in research results. *PLoS Biol, 16*(4), Article e2004299. https://doi.org/10.1371/journal.pbio.2004299

Plant, A. L., Locascio, L. E., May, W. E., & Gallagher, P. D. (2014). Improved reproducibility by assuring confidence in measurements in biomedical research. *Nature Methods, 11*(9), 895–898. https://doi.org/10.1038/nmeth.3076

Possolo, A. (2015). Simple guide for evaluating and expressing the uncertainty of NIST measurement results. *NIST Technical Note, 1900*. https://doi.org/10.6028/NIST.TN.1900

Potts, P. J. (2008). Glossary of analytical and metrological terms from the International Vocabulary of Metrology. *Geostandards and Geoanalytical Research*, *36*(3), 225–324. https://doi.org/10.1111/j.1751-908X.2011.00121.x

Reproducibility Initiative. (2014). Retrieved from http://validation.scienceexchange.com/#/about

Research Data Alliance Working Group on Persistent Identification. Retrieved from https://rd-alliance.org/groups/persistent-identification-instruments

Rosslein, M., Elliott, J. T., Salit, M., Petersen, E. J., Hirsch, C., Krug, H. F., & Wick, P. (2015). Use of cause-and-effect analysis to design a high-quality nanocytotoxicology assay. *Chemical Research in Toxicology, 28*(1), 21–30. https://doi.org/10.1021/tx500327y

Rouse, M. (2015). Fishbone diagram. *WhatIs.com*. Retrieved from https://whatis.techtarget.com/definition/fishbone-diagram

Stodden, V. M., S. (2014). Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Journal of Open Research Software*, *2*(1), e21.

http://doi.org/10.5334/jors.ay

Thompson, M., & Ellison, S. L. R. (2011). Dark uncertainty. *Accreditation and Quality Assurance*, *16*, Article 483. https://doi.org/10.1007/s00769-011-0803-0

Thompson, P. M. (2017, October 26). Tackling the reproducibility crisis requires universal standards. *Times Higher Education*. Retrieved from https://www.timeshighereducation.com/opinion/tackling-reproducibility-crisis-requires-universal-standards#survey-answer

Weir, K. (2015). A reproducibility crisis? *Monitor on Psychology*, *46*(9), 39. Retrieved from http://www.apa.org/monitor/2015/10/share-reproducibility.aspx

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

©2020 Anne L. Plant and Robert J. Hanisch. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

Footnotes

[1] The NASEM report’s definition of “replicability” is not equivalent to the VIM’s use of “replicate”, which is a sample that closely mimics another sample. ↩