Reproducibility and Replicability in Science, A Metrology Perspective

A Report to the National Academies of Sciences, Engineering and Medicine
Committee on Reproducibility and Replicability in Science

Anne L. Plant
Robert J. Hanisch
National Institute of Standards and Technology

15 June 2018

1 Introduction

The scope of this report is to highlight best practices that apply to research broadly, and specific areas of research that are particularly problematic. We will focus on tools and approaches for achieving measurement assurance, confidence in data and results, and the facility for sharing data.

The general nature of the problem. Concern about 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 [1], forensics [2], epidemiology [3], and psychology [4]. Failure to reproduce published results has been reported by researchers in chemistry, biology, physics and engineering, medicine, and earth and environmental sciences [5]. From the point of view of a national metrology institute, confidence in results from all fields of study are equally important and should be addressed thoroughly and systematically.

Reproducibility, Uncertainty, and Confidence

The role of reproducibility. Here 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 the various components of rigorous research may provide an alternative to a limited focus on reproducibility.

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 [6]. The International Vocabulary of Metrology[7] is commonly used by the international metrology community and provides definitions for many terms of interest to the issue of “reproducibility”. While the term has become something of a catch-phrase, “reproducibility” has a precise definition in measurement science. Table 1 lists a few of the terms in the VIM that describe the various aspects of a measurement process that relate to our discussion here.

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.	

Table 1. Some relevant terms and definitions that are consistent with the International Vocabulary of Metrology (VIM 2015). ‘Replicability’, a term that is often used in conjunction with ‘Reproducibility’, is not defined in the VIM.

There are many other sources of definitions in this space (e.g., [8]), but we point to the VIM because these definitions arise from measurement science, and have been developed over the course of decades through consensus by a large international community.

Reproducibility and the desire for confidence in research results. In addition to the occurrence of competing definitions associated with reproducibility, there are many caveats associated with the responses to the concern about reproducibility. Funding agencies, scientific journals, and private organizations have instituted checklists, requirements, and guidelines [9], [10], [11]. There have been a number of sponsored activities focused on demonstrating the reproducibility of previously published studies by other laboratories [12], [13]. Checklists have met with some resistance [14]. Some of the criticisms cited include 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. 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 [15], or worse, fraudulent. From a practical point of view, the effort to reproduce published studies can be prohibitively expensive and time consuming [16]. There are no easy answers for how to determine when the result of a complex study is sufficiently reproduced. It is not clear how to interpret the failure of an independent lab to reproduce another lab's results. "Who checks the checkers?" was a highly relevant question asked during an American Society for Cell Biology panel discussion on reproducibility. 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.

The complexities associated with interlaboratory reproducibility can be great, and when performed by metrology experts, interlaboratory studies follow a formal and systematic approach. 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 [17]), suggesting that reproducibility is not a sufficient indictor of confidence in a result. Mere reproducibility is insufficient to guarantee that a result of scientific inquiry indeed tracks the truth [18]. In addition, a failure to reproduce is often just the beginning of scientific discovery, and it may not be an indication that that any result is "right" or "wrong". Particularly in the case 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", and 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, by acknowledging and addressing sources of uncertainty in a research study. Thompson [19] goes further, suggesting that research methods should be reviewed and accredited as a prerequisite for publication of research in journals. Uncertainty in measurement and transparency of research methods are unifying principles of measurement science and the national metrology institutes.

The International Conventions of Metrology

Uncertainty in measurement is a unifying principle of measurement science and the national metrology institutes. The National Institute of Standards and Technology (NIST), which is the national metrology institute (NMI) of the United States, and its one hundred-plus sister laboratories in other countries quantify uncertainties as a way of qualifying measurements. This practice guarantees 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 [20]. 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[21]. As an example of their potential universality, the terms of the VIM have been adapted to provide a useful guide for geoscience research [22].

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 [23], the international metrology community lists a number of examples of sources of uncertainty (see Table 2).

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.

Table 2. Possible sources of uncertainty in a measurement (from the Guide to the Expression of Uncertainty in Measurement (GUM), Section 3.3.2 (JCGM, 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.

The sources of 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 will 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 a research setting, the formalism of calculating an expanded uncertainty is rarely necessary, but acknowledging and addressing sources of uncertainty is critical. Regardless of discipline, at each step of a scientific endeavor we should be able to identify the potential sources of 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 [24] (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. In addition, providing the 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.

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

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 [23] and other sources [27].
   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.

Table 3, reproduced from Plant et al. (2018) on "identifying, reporting, and mitigating sources of uncertainty in a research study."

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* [25]. This is a systematic way of charting all the experimental parameters 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 itemized in Table 3.

Reference materials. Instrument performance characterization and experimental protocol evaluation are aided by the use of Reference Materials and Standard Reference Materials® (SRM). SRMs are the most highly characterized reference materials produced by NIST. RMs and SRM 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. Such a material is referred to as a NIST Traceable Reference Material™**.**

Calibration services. NIST provides the highest order of calibration services for instruments and devices available in the United States. 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 that provide to 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, but 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 [26]. Cell seeding density and cell washing steps were two variables that were systematically explored for their effect and yielded the knowledge that it was important to specify these protocol details. As a result of the analysis, a *series of in process controls* were run with every measurement. The results of the controls 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. Additional control experiments that were reported included small tandem repeat analysis of the cell lines used in the different laboratories, and analysis of nanoparticle aggregation. 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, experimental science laboratories that participate in formal inter-laboratory studies [27]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 one hundred 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, but subject to expert review and assessment of measurement practices and uncertainty characterization. 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 serve as an exemplar of the kind of processes that, if adopted more widely, would improve confidence in research data generally.

4 Thorny 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. A few example are presented below.

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)[1] 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 by different laboratories. 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, and that there are other regions where it may be impossible to establish a high level of confidence. Putting a numerical value on concordance under these circumstances is challenging.

Complexity of research studies and measurement systems. Part of the challenge in genome sequencing, and 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, determining 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

[1] http://jimb.stanford.edu/giab/

and compare the results from, the real answer—the ground truth sequence—isn't known. DNA sequencing is certainly not the only example of this dilemma.

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 is unlikely in a research environment and 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 small tandem repeat (STR) sequences. Above 75% concordance in STR sequences identified was determined to be sufficient for identification [28].

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 have similar, but not identical, behavior. Indeed, one could argue that climate studies and predictions are of this nature, given that it is impossible to run a controlled experiment.

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. How many research papers are there that include the throw-away line "the data were reduced in the usual manner"? Meaning, of course, that no one bothered to record the various input parameters and options. As noted by Stodden and Miguez [29], documenting what software was used and sharing code are essential practices for assuring reproducible and reusable research. Increasingly we are seeing the publication and sharing of data and processing steps in, for example, Jupyter[2] notebooks, and also the registration of software packages and source codes in shared indexes (e.g., the Astrophysics Source Code Library[3] or the NIST Materials Resource Registry[4] [30]. In fact, the Materials Resource Registry indexes both data and software, treating the latter as a special type of data.

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 rigor* to demonstrate confidence in these components of a study by providing supporting evidence. At the very least, researchers should share data and software, including source code. Outside of computer science, the unreliability of software is often underappreciated. Rigorous testing of

[2] http://jupyter.org
[3] http://ascl.net
[4] https://materials.registry.nist.gov/

software should be performed, as it has long been understood that numerical software has reliability challenges [31]. Testing the ability to produce the same results from computer code running in different machines under different operating systems and with the same inputs provides an indication that the results are generalizable beyond a particular computing environment [32], [33].

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 theorists, 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 [34], and NIST’s Thermodynamics Data Center has long employed partially automated data quality assessment tools [35]. 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,[5] 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. Unambiguous meanings and context in metadata labels would enable searching and discovery of similar and dissimilar experimental protocol details. 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 range of applicability, reliability, and uncertainty.

[5] https://rd-alliance.org/groups/persistent-identification-instruments

How much reporting is enough? Irreproducibility can be an important and positive factor in advancing science. 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 1930's 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[6]. Often, irreproducibility is the result of failure to identify and control major sources of variability. Medical meta-analysis [36] often will increase the state of knowledge about the performance of a therapy even when different studies produce inconsistent results. Irreproducibility can indicate that some parameter that has not been controlled is an important source of uncertainty [37]. In biomedical research, there can be so many uncontrolled and hidden variables that there is a high likelihood that experiments preformed 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 report every experimental variable, protocol nuance, and instrument parameter. 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 make it more feasible to fully describe our research studies.

6 Discipline-Specific Considerations

The importance of reproducibility relative to other aspects of the scientific process can be different for different scientific disciplines. However, 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.

Astronomy

In the report of the NSF-sponsored Workshop on Robustness, Reliability, and Reproducibility in Scientific Research,[7] a diagram sketched by Roger Peng is reproduced in which various research disciplines are categorized as whether they are support by strong basic theory or not, and whether they have a tradition of controlled experiments. Astronomy is shown to be supported by theory—true, but as not having a tradition of controlled experiments.

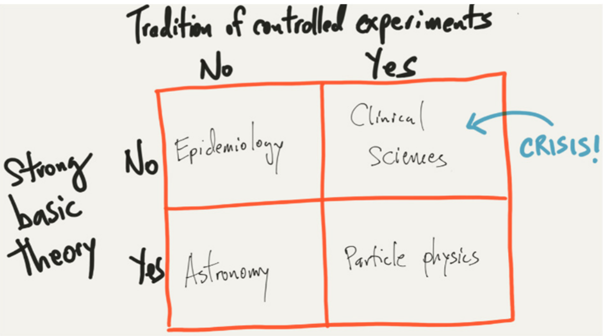


Figure 1. Roger Peng's characterization of various research disciplines. From Weitz (2017).

[6] https://www.cfa.harvard.edu/~dfabricant/huchra/hubble/

[7] http://www.mrsec.harvard.edu/2017NSFReliability/include/NSF_Workshop_Robustness.Reliability.Reproducibility.Report.pdf

With the advent of large-scale digital sky surveys and routine pipeline-based calibrations that produce science-ready data products, this does not seem to be an accurate characterization. Hanisch has worked in astronomy for more than thirty years, co-leading the development of the Multi-Mission Archive at Space Telescope and directing the US Virtual Astronomical Observatory. The vast majority of astronomical research data is open (often after a nominal proprietary period such as 12 months), leading 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).[8] The Sloan Digital Sky Survey 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.[9] 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) reproducibility problems are rare. Where they do exist, as in the Hubble constant studies mentioned earlier, they result from incomplete information about the phenomenon being measured.

Physics

The report from the NSF-sponsored workshop mentioned above[10] focused primarily on the research disciplines within NSF's Directorate for Mathematical and Physical Sciences. Its primary conclusion was "that, for the scientific disciplines within MPS, the science process works, and that there is not a 'crisis' of robustness, reliability, and reproducibility." We believe, however, that the situation is more complex, and that this report's characterization of some disciplines as "mature" and others, by contrast, as "immature" casts an unnecessarily pejorative tone.[11] 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 that neutrinos produced at CERN travelled faster than the speed of light [38]. 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 and where hidden variables lead to challenges in reproducibility.

[8] https://archive.stsci.edu/hst/bibliography/pubstat.html

[9] Astrophysics Data System search, 11 May 2018, https://ui.adsabs.harvard.edu/

[10] http://www.mrsec.harvard.edu/2017NSFReliability/include/NSF_Workshop_Robustness.Reliability.Reproducibility.Report.pdf

[11] One of us, Hanisch, was a participant in the workshop and co-edited the report. He tried (unsuccessfully!) to delete the "mature" characterizations from the document.

Materials Science

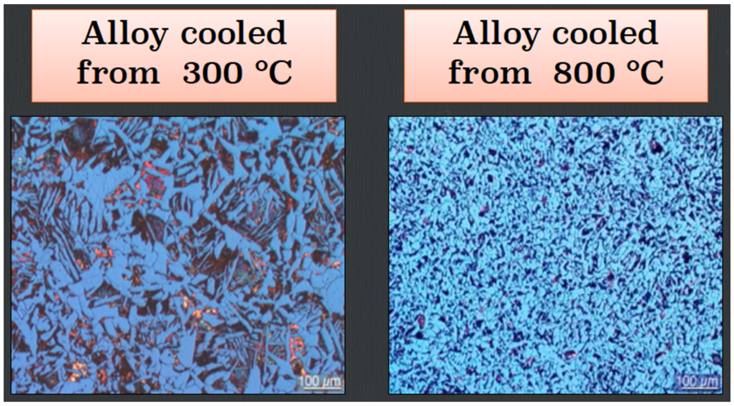


Figure 1. Grain structure in alloys depends strongly on cooling processes. (Credit: J. Warren, NIST.)

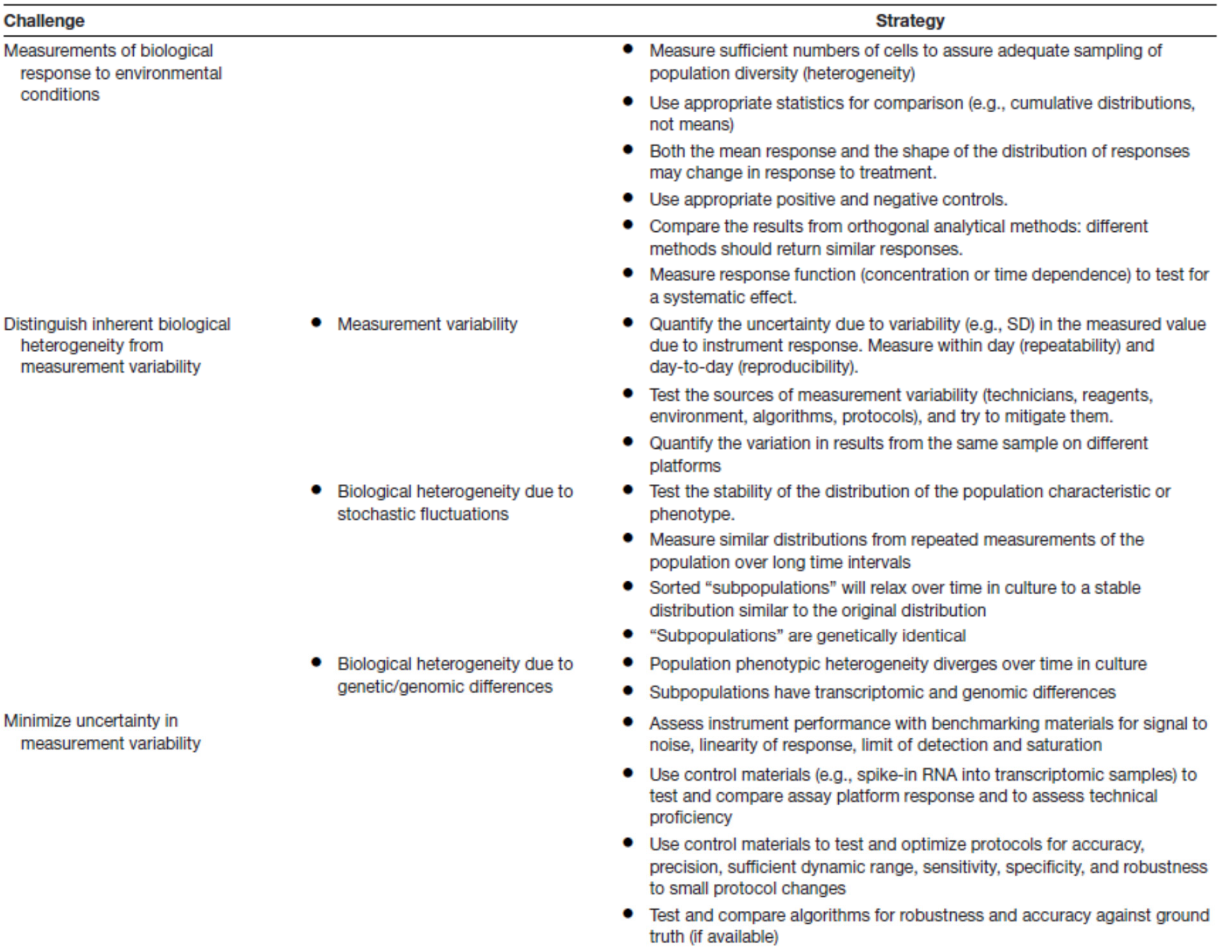
Materials science is a major area of research at NIST, and NIST manages the national Materials Genome Initiative (MGI),[12] whose goal is to accelerate the development of new materials at lower cost through better integration of computer simulation and experimentation. There are significant challenges 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. If the growth and processing history of a material is not fully documented, if unknown (and hence unmeasured) effects impact properties, or if significant instrumental parameters are hidden in proprietary binary data formats, reproducibility suffers.

Biology

For highly complicated studies that involve a very large number of parameters such as are 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 use to have sufficient information and facility to compare exactly which aspects of which steps in the processes were 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, 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, lead to inherent heterogeneity in biological responses. 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 appropriate

[12] https://www.mgi.gov/

metric since the biological response function is often not well-described by a Gaussian distribution. Table 4 is taken from Keating et al. [39], and articulates some challenges and strategies in single cell experiments for distinguishing measurement uncertainty from biological variability.



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 the shape of the distribution of responses may change in response to treatment. • Use appropriate positive and negative controls. • Compare the 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 stochastic fluctuations	• Test the stability of the distribution 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 algorithms for robustness and accuracy against ground truth (if available)

Figure 4. Distinguishing measurement uncertainty from biological variability in a single cell assay (from [39]).

While techniques like design of experiment can be used to assess interactions between multiple parameters 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 datasets will be instrumental to recognizing the patterns of events in complex systems and for developing the understanding of fundamental principles for predicting outcomes in complex systems. More than ever, we must have confidence in the data that will be available for development of models of such complex systems.

7 Qualifying and Characterizing Measurement Systems

Table 5 presented by Plant et al. [40] was developed based on criteria provided by the U.S. Food and Drug Administration for qualifying assays that are used to characterize a regulated biological product. This measurement elements in the Table are criteria that, when identified, help to provide confidence about the assay and the results of the assay. Achieving the knowledge of these measurement elements requires a high level of understanding of, and experience with the assay. Reporting these characteristics for an assay provides two advantages. One advantage is for the user of the assay. Particularly when a regulated biological product is being evaluated, it is critical for the manufacturer to have clear expectations about performance of the characterization assays. It is critical that when an assay provides an answer, that there is a high level of confidence that the assay result is an accurate assessment of the product. Many regulated biological products result from precious samples and expensive and lengthy manufacturing processes; it is critical that if an assay indicates ambiguity it is clear that that ambiguity arises from the product, and that the veracity of the assay is non-ambiguous. Clearly another important use of these criteria is to establish confidence for the regulatory process.

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 interlaboratory 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.

Table 5. Reproduction of Table 1 from Plant et al. (2014) showing key elements of a good measurement.

8 May 2018 Workshop

On May 1–3, 2018 a workshop[13] entitled "Improving Reproducibility in Research: The Role of Measurement Science," was hosted by the National Physical Laboratory, Teddington, UK and was co-organized by NPL, NIST, and several other national metrology institutes (NMIs). From the workshop flyer,

> "The goal of this workshop is to bring together experts from the measurement and wider research communities to understand the issues and to explore how good measurement practice and principles can foster confidence in research findings; including how we can tackle the challenge posed by increasing data volumes in both industry and research."

Approximately 80 individuals registered for the workshop, representing academia, industry, and government (NMIs in particular). The format of the workshop included plenary presentations, panel discussions, breakout sessions, and a road-mapping exercise. The ultimate outcome of the workshop will be a report with recommendations for actions the network of NMIs, working in collaboration with the BIPM, can take to help improve reproducibility and confidence in research.

The road-mapping exercise exposed a potential list of areas where the metrology community can hope to make an impact. We note that these activities have not yet been prioritized nor endorsed by the individual NMIs.

Intercomparisons and Replication Studies

- Noted importance of key comparisons and other interlaboratory studies specifically aimed at measurement of the same measurand.
- Should aim to assure that all outputs of research studies are replicable through machine-actionable data and metadata (such as Jupyter notebooks).
- Should aim to have data from all measurements to be openly available with calibration certificates.
- Intercomparisons require consistent vocabularies and ontologies in order to be interoperable.

Repeatability and Reproducibility

- Some fields such as the pharmaceutical industry require comparisons prior to approval of a new drug.
- Instrumentation needs to have readily accessible metadata for all information affecting data and measurements, preferably in open, non-proprietary formats.
- The NMIs should lead by example, demonstrating best measurement practices internally and sharing these with the broader research community.
- Increase automation of data acquisition so as to minimize potential for human error

[13] http://www.npl.co.uk/improving-reproducibility-in-research

- Aim for automatic capture of the research process (workflow), through to publication of machine-actionable research articles.
- Reward scientists who produce reproducible, re-usable research data and software.

Training

- Noted that training in the principles of metrology, uncertainty characterization, statistical methods, and machine learning is largely lacking in the university curriculum.
- NMIs could consider assisting in developing best-practice guidelines, provide open source datasets for training and demonstration of proficiency, and create a universal platform for access to training materials; collaborate with data science training programs sponsored by CODATA.
- Metrology hackathons for uncertainty estimation in AI and machine learning.
- Establish Software Carpentry[14]-like program for exposure to sound measurement methodologies.

International Standards for Data

- Use, adapt, and adopt; numerous metadata standards exist. Where they are lacking first consider extending existing ones.
- Ideally there would be fewer standards, but each with higher adoption rates.
- Need to assure that standards incorporate proper metrology (e.g., unambiguous expression of units of measure).
- Develop comprehensive directory of relevant standards and their purpose/scope.

Reference Materials, Reference Data

- There is a major gap in reference materials in the bioclinical and materials science research areas.
- Broader use of reference materials and reference data would improve research reproducibility and confidence in measurement.

Traceability

- Consider establishing a Consultative Committee on Data under the auspices of BIPM
- Establish measurement standards and best practices for research areas that must deal with large numbers of hidden variables, sparsely sampled data, etc.
- Require machine-readable provenance for research data.
- Define framework for uncertainty, reliability, and provenance for AI and machine learning.

[14] https://software-carpentry.org/

9 Conclusions

We hope to have demonstrated that concepts and practices in the metrology community—if brought to bear on the practice of scientific research more broadly—could have a profound effect on the quality of research outputs and thereby increase confidence in the conclusions drawn. Some of the most important steps to be taken 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, 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 principles to research data broadly
- 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

We emphasize that non-reproducible research is not necessarily indicative of bad science, and that disagreement between experiments often arises because not all aspects affecting the measurement are known. Indeed, it is through such inconsistencies that science advances.

References

1. Peng RD: **Reproducible research in computational science**. *Science* 2011, **334**(6060):1226-1227.
2. **Strengthening Forensic Science in the United States: A Path Forward**. In. Edited by Council NR: Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council; 2009.
3. Ioannidis JP, Bernstein J, Boffetta P, Danesh J, Dolan S, Hartge P, Hunter D, Inskip P, Jarvelin MR, Little J *et al*: **A network of investigator networks in human genome epidemiology**. *Am J Epidemiol* 2005, **162**(4):302-304.
4. Open Science C: **PSYCHOLOGY. Estimating the reproducibility of psychological science**. *Science* 2015, **349**(6251):aac4716.
5. Baker M: **1,500 scientists lift the lid on reproducibility**. *Nature* 2016, **533**(7604):452-454.

6. Possolo A: **Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results** *NIST Technical Note* 2015, **1900**.

7. **International Vocabulary of Metrology – Basic and General Concepts and Associated Terms**. In.: BIPM; 2012.

8. Pellizzari ED, Lohr KN, Blatecky AR, Creel DR: **Reproducibility : a primer on semantics and implications for research**. Research Triangle Park, NC: RTI Press; 2017.

9. Collins FS, Tabak LA: **Policy: NIH plans to enhance reproducibility**. *Nature* 2014, **505**(7485):612-613.

10. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G *et al*: **SCIENTIFIC STANDARDS. Promoting an open research culture**. *Science* 2015, **348**(6242):1422-1425.

11. Boisvert RF: **Incentivizing Reproducibility**. *Communications fo the ACM* 2016, **59**(10):5.

12. **Reproducibility Initiative** [http://validation.scienceexchange.com/#/about]

13. Weir K: **A reproducibility crisis?** *Monitor on Psychology* 2015, **46**(9):39.

14. Baker M: **US societies push back against NIH reproducibility guidelines**. *Nature* 2015.

15. Lash TL: **Declining the Transparency and Openness Promotion Guidelines**. *Epidemiology* 2015, **26**(6):779-780.

16. Maher B, : **Cancer reproducibility project scales back ambitions**. *Nature* 2015.

17. Mallove EF: **Fire from ice : searching for the truth behind the cold fusion furor**. New York, N.Y.: J. Wiley; 1991.

18. Roush S: **Tracking Truth: Knowledge, Evidence and Science**. Oxford: Oxford University Press; 2006.

19. Thompson PM: **Tackling the reproducibility crisis requires universal standards**. In: *Times Higher Education.* 2017.

20. **Brief history of the SI** [http://www.bipm.org/en/measurement-units/history-si/]

21. **Evaluation of measurement data — Guide to the expression of uncertainty in measurement**. In. Edited by Metrology JCfGi: BIPM; 2008.

22. Potts PJ: **Glossary of Analytical and Metrological Terms from the International Vocabulary of Metrology** *Geostandards and Geoanalytical Research* 2008, **36**(3):225-324.

23. **Evaluation of measurement data — Guide to the expression of uncertainty in measurement**. In., vol. JCGM100:2008 Joint committee for guides in metrology; 2008.

24. Plant AL, Becker CA, Hanisch RJ, Boisvert RF, Possolo AM, Elliott JT: **How measurement science can improve confidence in research results**. *PLoS Biol* 2018, **16**(4):e2004299.

25. Rouse M: **Fishbone diagram**. In: *WhatIscom.* TechTarget; 2015.

26. Rosslein M, Elliott JT, Salit M, Petersen EJ, Hirsch C, Krug HF, Wick P: **Use of Cause-and-Effect Analysis to Design a High-Quality Nanocytotoxicology Assay**. *Chem Res Toxicol* 2015, **28**(1):21-30.

27. Hibbert DB: **Quality Assurance for the Analytical Chemistry Laboratory**: Oxford University Press 2007.

28. American Type Culture Collection Standards Development Organization Workgroup ASN: **Cell line misidentification: the beginning of the end**. *Nat Rev Cancer* 2010, **10**(6):441-448.

29. Stodden VM, S.: **Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research**. *Journal of Open Research Software* 2014, **2**(1):e21.

30. Becker CA, Dima A, Plante RL, Youssef S, Medina-Smith A, Bartolo LM, Hanisch RJ, Warren JA, Brady MC: **Development of the NIST Materials Resource Registry as a means to advertise, find, and use materials-related resources**. *Materials Science and Technology Society* 2017.

31. Boisvert RF, International Federation for Information Processing.: **Quality of numerical software : assessment and enhancement**, 1st edn. London ; New York: Published by Chapman & Hall on behalf of the International Federation for Information Processing; 1997.

32. Mytkowicz T, Diwan A, Hauswirth M, Sweeney PF: **Producing Wrong Data Without Doing Anything Obviously Wrong!** *Acm Sigplan Notices* 2009, **44**(3):265-276.

33. Blackburn SMD, A.; Hauswirth, M.; Sweeney, P.F.; Amaral, J.N.; Brecht, T.; Bulej, L.; Click, C.; Eeckhout, L.;Fishchmeister, S.; Frampton, D.; Hendren, L.J.; Hind, M.; Hosking, A.L.; Johnes, R.E.; Kalibera, T.; Keynes, N.; Nystrom, N.; Zeller, A.: **The Truth, The Whole Truth, and Nothing But the Truth: A Pragmatic Guide to Assessing Empirical Evaluations**. *ACM Transactions on Programming Languages and Systems* 2016, **38**(4).

34. McIntosh L, et al.: **Repeat: A Framework to Assess Empirical Reproducibility in Biomedical Research**. In: *OSFPreprints.* 2017.

35. Frenkel M, Chirico RD, Diky V, Yan XJ, Dong Q, Muzny C: **ThermoData engine (TDE): Software implementation of the dynamic data evaluation concept**. *J Chem Inf Model* 2005, **45**(4):816-838.

36. Cooper H, Patall EA: **The relative benefits of meta-analysis conducted with individual participant data versus aggregated data**. *Psychol Methods* 2009, **14**(2):165-176.

37. Thompson ME, S.L.R.: **Dark uncertainty**. *Accreditation and Quality Assurance* 2011, **16**:483-487.

38. Brumfiel G: **Neutrinos not faster than light**. *Nature* 2012.

39. Keating SM, Taylor DL, Plant AL, Litwack ED, Kuhn P, Greenspan EJ, Hartshorn CM, Sigman CC, Kelloff GJ, Chang DD *et al*: **Opportunities and Challenges in Implementation of Multiparameter Single Cell Analysis Platforms for Clinical Translation**. *Clin Transl Sci* 2018, **11**(3):267-276.

40. Plant AL, Locascio LE, May WE, Gallagher PD: **Improved reproducibility by assuring confidence in measurements in biomedical research**. *Nat Methods* 2014, **11**(9):895-898.