Volume 124, Article No. 124024 (2019) https://doi.org/10.6028/jres.124.024
Journal of Research of the National Institute of Standards and Technology

Improving Reproducibility in Research: The Role of Measurement Science

Robert J. Hanisch[1], Ian S. Gilmore[2], and Anne L. Plant[1]

[1]National Institute of Standards and Technology,
Gaithersburg, MD 20899, USA

[2]National Physical Laboratory,
Teddington, TW11 0LW, United Kingdom

robert.hanisch@nist.gov
ian.gilmore@npl.co.uk
anne.plant@nist.gov

Summary:

- We report on a workshop held 1–3 May 2018 at the National Physical Laboratory, Teddington, U.K., in which the focus was how the world's national metrology institutes might help to address the challenges of reproducibility of research.
- The workshop brought together experts from the measurement and wider research communities in physical sciences, data analytics, life sciences, engineering, and geological science. The workshop involved 63 participants from metrology laboratories (38), academia (16), industry (5), funding agencies (2), and publishers (2). The participants came from the U.K., the United States, Korea, France, Germany, Australia, Bosnia and Herzegovina, Canada, Turkey, and Singapore.
- Topics explored how good measurement practice and principles could foster confidence in research findings and how to manage the challenges of increasing volume of data in both industry and research.

Key words: confidence; replicability; uncertainty.

Accepted: September 5, 2019

Published: September 18, 2019

https://doi.org/10.6028/jres.124.024

1. Motivation and Scope

Much has been written in the press recently suggesting that there is a "reproducibility crisis" in scientific research. This stems from well-publicized papers such as those by Brian Nosek *et al.* of the Center for Open Science noting the difficulties in replicating research findings in a number of papers in psychology journals [1], and books such as *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hopes, and Wastes Billions* by Richard Harris [2]. Unfortunately, these publications sensationalize the problems—some of which are real and some of which are not—and threaten to undermine public confidence in scientific research generally [3].

In mid-2017, Anne Plant and Robert Hanisch conceived of a workshop that would address, from a measurement science perspective, a reasoned and rational narrative of the problems underlying repeatability, replicability, and reproducibility. At the same time, staff at the National Physical Laboratory (NPL), Teddington, U.K., began engaging in this topic, motivated in particular by the challenges of dealing with big data. These initiatives came together at the October 2017 meeting of National Metrology Institute (NMI) directors at the Bureau International des Poids et Mesures (BIPM), where Joern Stenger (Physikalisch-Technische Bundesanstalt, PTB) led a session titled "Advanced Manufacturing, Digitization, and the Internet of Things." This addressed the question of the role of metrology in the world of "Big

How to cite this article:
Hanisch RJ, Gilmore IS, Plant, AL (2019) Improving Reproducibility in Research: The Role of Measurement Science.*J Res Natl Inst Stan* 124:124024. https://doi.org/10.6028/jres.124.024

Data," where increasing digitization and increasing volumes of data can lead to issues of confidence in industry and in research reproducibility. During the discussion, Hanisch and Martyn Sené (NPL) indicated that both their organizations were considering organizing workshops to look at the role of metrology and the NMI community in related areas. Subsequently, the two organizations joined forces in planning the workshop held at NPL on 1–3 May 2018, bringing in representatives from a number of other NMIs and related organizations on the Organizing Committee (see Acknowledgments).[1, 2]

2. Format

The workshop consisted of plenary sessions with invited talks, moderated panel discussions, topical breakout discussions, and a lightly structured road-mapping exercise. There was a very high level of engagement throughout the workshop.

Exploration of the Problem (Tuesday 1 May 2018)—Chair: Robert Hanisch

Pete Thompson	Welcome and Opening Remarks
Mark Thomson	Reproducibility Challenges in U.K. Science
Eric Loken	Measurement, Statistical Intuitions, and the Replication Crisis in Science
Barend Mons	FAIRSharing's Role in Addressing Confidence in Research
Hilary Hanahoe	Research Data Alliance Activities
Geoffrey Boulton	CODATA/ISC Activities (Committee on Data for Science and Technology/International Council for Science)
Simon Cox	Precision in Nomenclature for Transfer of Observation Data

Panel Discussion: Domain-Based Challenges

Anne Plant	Biosciences
Jim Warren	The Materials Genome Initiative and Reproducibility
Stephen Ellison	Reproducibility in Chemical Research
Michael Hildreth	Reproducibility in Physics

Panel Discussion: Technique-Based Challenges

Ian Gilmore	Mass Spectrometry
June Lau	Reproducibility Challenges in Electron Microscopy
Steve Collins	Measurement Reproducibility Issues: Synchrotron and Neutron Facilities

[1] This publication is intended to capture external perspectives related to National Institute of Standards and Technology (NIST) standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and other organizations. This report was prepared as an account of a workshop; it is intended to document external perspectives and does not represent official NIST positions. The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies of NIST or the U.S. government.

[2] Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

John Elliott	Biosciences
J.T. Janssen	Quantum
	Panel Discussion: Data Life Cycle
Keith Jeffery	Metadata and Data Models
Robert Hanisch	Data Management and Provenance
John Henry Scott	Algorithms, Software, and Data
Eva Campo	Data Preservation and Reuse Goals, a National Science Foundation (NSF) Perspective
Helen Glaves	Machine-Actionable Data Management Plans
Martyn Sené	Closing Remarks for Day 1

NMI Resources (Wednesday 2 May 2018)—Chair: Ian Gilmore

Tony Hey	Reproducibility in Computation and Artificial Intelligence (AI)
Owen Sansom	Reproducibility in Cancer Research
Antonio Possolo	Trustworthy Measurement for Reproducible Research
Martin Milton	Reproducibility Is Our Business
	Panel Discussion: Reference Data, Reference Materials, and Intercomparisons
Robert Hanisch	NIST SRD (National Institute of Standards and Technology Standard Reference Data)
Hyun Kyoon Lim	KRISS SRD (Korea Research Institute of Standards and Science Standard Reference Data)
Ray Plante	Metrology Resource Registry
Jeanita Pritchett	Reference Materials
Graham Sims	Versailles Project on Advanced Materials and Standards (VAMAS) and Prenormative Data
Stuart Chalk	SI for Dummies[3]
	Panel Discussion: Measurement Practice
Jeanita Pritchett	Metrology Education and Accreditation
Jan Jensen	Achieving Process Understanding When Directing Cellular Differentiation
Alistair Forbes	Trusted Data, Trusted Models, Trusted Algorithms, Trusted Software
Colin Longstaff	Measurement Practices in Biology
Sir Jim Smith	A Wellcome Approach to Reproducibility

[3] SI is the International System of Units.

Breakout Session 1

AI and Machine Learning

Uncertainty Quantification

Statistics and P-Factors, Bayesian Methods

Breakout Session 2

Tools to Improve Confidence in Measurement

The FAIR Principles and Reproducibility

Reports Out

Broader Implications of Reproducibility (Thursday 3 May 2018)—Chair: J. T. Janssen

Leslie McIntosh	Standards for Research Reproducibility
Natalie De Souza	The Role of Journals in Promoting Research Standards and Reproducibility
Anne Plant	Summary Observations

Copies of the workshop presentations are available at http://www.npl.co.uk/insights/improving-reproducibility-in-research.

3. Summaries of Invited Talks

Mark Thomson, executive chair of the U.K. Science and Technology Facilities Council (STFC), gave the opening talk on “Reproducibility Challenges and U.K. Science.” He set the STFC in context with the newly formed U.K. Research and Innovation (UKRI) organization for funding research and innovation in the U.K. UKRI brings together seven U.K. research councils, Innovate U.K., and Research England. This provides a unified voice for the U.K.’s research and innovation system to create a smoother pathway for innovation. An immediate priority is engagement with stakeholders to create a roadmap to reach 2.4 % of gross domestic product (GDP) investment in research and development (R&D) by 2027. The UKRI goal is for the U.K. to continue to be a world leader in research and innovation. Having reliable published scientific results is critical to this goal. Thomson reviewed some of the literature on the reproducibility challenge, including a *Nature* survey of 1576 researchers [4, 5] illustrating the threats to reproducible science. These threats include failure to control for bias, low statistical power, poor quality control, p-hacking (selection of research results that support a significant outcome), publication bias, and hypothesizing after the results are known (HARKing). It is often thought that the issues are limited to particular scientific fields, but the *Nature* survey showed that issues are present in all scientific disciplines. He gave examples from his own field of particle physics, for example, showing how, over time, experimental measurement of the neutron lifetime converged asymptotically owing to unconscious bias towards previous measurements. This community has taken the approach of “blind analysis” to help reduce bias. The *Nature* survey asked “What factors could boost reproducibility?” Better understanding of statistics, better mentoring, more robust design, and more within-laboratory validation were the highest ranked responses. He noted that these issues directly map onto the goals of metrology, for example, developing robust methodologies and better understanding and reporting of statistics and uncertainty. He highlighted recent papers from NIST [6, 7] and NPL [8] that provide a perspective on how the metrology community can help. In summary, he highlighted that reliable, reproducible, and robust experimental data and interpretation underpin the scientific method and that they inform the flow of resources and policy. The evidence shows that there is a reproducibility challenge, and, unfortunately, some incentives do not necessarily encourage good practice. The whole scientific community has a role to play in addressing the issue, and the metrology community is important to this debate.

Erik Loken (University of Connecticut) spoke about "Measurement, Statistical Intuitions, and the Replication Crisis in Science." Citing one of the major reports about unreliable research, he noted that there are substantial challenges in reproducibility in fields such as genetics, neuroscience, medicine, psychology, nutrition, and education. The underlying causes are complex, including promotion and tenure processes that encourage a focus on the quantity of publications and citation indices. However, in his talk, Loken focused on two core problems: unintentional p-hacking ("sincere research," as opposed to the "insincere" practice of intentional p-hacking) and weak statistical intuition. Statistical intuition includes topics such as the null hypothesis testing framework, overestimated effect sizes, whether measurement error attenuates effects, and understanding the covariance structures being estimated. P-test hacking, or basically cherry-picking or filtering data to reach a preferred outcome, is clearly a problem in some fields. One solution to this that is currently being tested is pre-registration of analysis plans (see the Center for Open Science, Open Science Framework platform, for instance).[4]

Barend Mons (GO FAIR, University of Leiden) gave an overview of the FAIR (Findable, Accessible, Interoperable, and Reusable) data framework and its current instantiation in the GO FAIR[5] initiative in Europe. In particular, he focused on "implementation networks," where research disciplines come together to support and build FAIR data repositories and services. There are already 30 implementation networks in various stages of maturity, and the intent here was to establish an implementation network amongst the metrology institutes organized under the BIPM. To date, the metrology implementation network remains in a formative state, though through organizations such as NIST and NPL, we intend to provide and promote FAIR data services.

Hilary Hanahoe (secretary general, Research Data Alliance) discussed the role that the Research Data Alliance (RDA) is playing in improving the access to and interoperability of research data from across the disciplinary spectrum. The RDA vision is "building the social and technical bridges to enable open data sharing." With over 8000 individual members from 120 countries, organized into more than 100 working groups and interest groups, RDA brings together data experts from diverse fields to build consensus around optimal data-sharing infrastructure and policies. RDA Working Groups are formed to develop and implement data infrastructure over a period of 12–18 months and have Recommendations as formal outputs. Interest Groups can be active for an indefinite period of time and focus on solving specific data-sharing problems, often initiating a Working Group to carry out the technical activities. There is currently an Interest Group specifically focused on reproducibility.[6]

Geoffrey Boulton (president, CODATA and University of Edinburgh) gave an update on the International Science Council (ISC) and the Committee on Data for Science and Technology (CODATA). The purpose of CODATA is to improve the quality, reliability, management, accessibility, and use of data of importance in all fields of research. The committee was established in the 1960s with a priority on fundamental constants of physics and chemistry. Over the decades, these priorities have developed to respond to changing needs, and now the focus is on the digital revolution and open science. This focus

[4] https://osf.io
[5] https://www.go-fair.org
[6] https://rd-alliance.org/groups/reproducibility-ig.html

recognizes today's vast data streams, diversity, and computational capacity. The data infrastructure allows instantaneous access, anywhere, anytime, at low costs. CODATA has three strategic priority areas: (1) principles, policies, and practice; (2) advancing the frontiers of data science (including the *Data Science Journal*); and (3) mobilizing data capacity. Further details are available in the *CODATA Prospectus: Strategy and Achievement, 2015–2017* [9]. Dr. Boulton highlighted the fact that shared, standardized terminologies are increasingly important to enable wider integration of data, especially in interdisciplinary science.

Simon Cox (Commonwealth Scientific and Industrial Organisation (CSIRO, Australia) discussed precision in nomenclature for transfer of observation data. He highlighted the need for a standard vocabulary for observational data that would support significant complexity, including the protocols for the observation, the sensors used, the results, the result time, and also information on the timescale of the phenomena under observation. The World Wide Web Consortium (W3C) is an international community that develops open standards to ensure the long-term growth of the Web. They have published the Semantic Sensor Network (SSN) ontology for describing sensors and their observations, the procedures involved, the features of interest studied, the samples used to do so, and the observed properties, as well as actuators.[7] This has now become an accepted standard in practice. However, a standard for machine-readable units of measure is still lacking. The Unified Code for Units of Measure (UCUM)[8] is a code system intended to include all units of measure being contemporarily used in international science, engineering, and business. The purpose is to facilitate unambiguous electronic communication of quantities together with their units. The focus is on electronic communication, as opposed to communication between humans. Typical applications of the UCUM are electronic data interchange (EDI) protocols, but there is nothing that prevents it from being used in other types of machine communication.

Tony Hey (U.K. Science and Technology Facilities Council, Rutherford Appleton Laboratory) examined reproducibility challenges in the areas of computation and artificial intelligence (AI). Computational science is the so-called "third paradigm," the first being experiment, the second theory, and the fourth being data-intensive science [10]. It had been noted already in 2012 that computational experiments, *i.e.*, simulations, frequently do not have documented workflows, lists of software used and dependencies, or details about the computational hardware and compiler options. For complex codes running on massively parallel systems, numerical round-off errors can be greatly magnified. Similarly, different codes attempting to address the same physical problem but using different algorithms are unlikely to obtain numerical agreement. For AI and machine learning, it is important to have well-characterized training and test data sets, though experiments have shown that deep learning networks can be easily fooled when "adversarial noise" is added to images. Hey gave several examples of machine learning applications in research: the Dark Energy Survey in astronomy, for galaxy/star identification, and the Large Hadron Collider Compact Muon Solenoid experiment, for identification of subatomic particles. Hey concluded by noting the need for training of data scientists and the incorporation of sound data science practices in the research community generally.

Antonio Possolo (fellow, Statistical Engineering Division, Information Technology Laboratory, NIST) spoke on trustworthy measurement that tracks the truth sufficiently closely, with assuredly high confidence. A trustworthy measurement (1) is traceable and well calibrated, (2) has an uncertainty evaluation that is fit for the purpose, and (3) is validated through comparability, consistency, or consilience (convergence of evidence). To "track the truth," the measurement needs to be traceable and calibrated, and the method must be validated with "check standards." The term "sufficiently closely" was explained as a measurement

[7] https://www.w3.org/TR/vocab-ssn/
[8] http://unitsofmeasure.org/trac/

uncertainty that is sufficiently small such that the measured value is an effective proxy for the true value and such that the result is fit-for-purpose. The final term “assured high confidence” was characterized as there being a high credence in the true value lying within the margin of reported uncertainty surrounding the measured value and that mutual agreement of measurement results can be independently achieved using different methods. It was further noted that the reported uncertainties should be corroborated in an intercomparison (*i.e.*, there is no significant dark [unknown or unaccounted for] uncertainty). To improve reproducibility, Possolo concluded that researchers should rely on trustworthy measurements and design research for consilience, so that research conclusions are substantiated using multiple independent, essentially different experimental methods.

Martin Milton (director of BIPM, or the International Bureau of Weights and Measures) opened his presentation by asserting that “reproducibility is our business.” The BIPM is the international organization that coordinates matters related to metrology (measurement science), and it was established by the Metre Convention of 1875. Over 100 countries participate in the Metre Convention and work under the auspices of BIPM as member states or associate states. Milton noted that the objectives of metrology are to assure stable, comparable, and coherent measurements. Within the field of metrology, “reproducibility,” “replicability,” and “repeatability” have specific meanings [11]. Reproducibility denotes the closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement. This goes beyond “repeatability” (can I get consistent results from my experiment?) and “replicability” (can someone else get consistent results in duplicating my experiment?) and thus is the strongest test of reliability of a measurement. The key concept in metrology is “traceability,” *i.e.*, that there is a documented, unbroken chain of calibrations linking a measurement to the fundamental physical constants. Such measurements are furthermore characterized by “uncertainty,” a parameter that describes the dispersion of the quantity values being attributed to a measurand. BIPM maintains a Key Comparisons Database[9] in which the practice of metrology is instantiated. Various metrology institutes around the world make precise measurements using different techniques and share their results and uncertainties. These comparisons are a direct indication of the reproducibility of a measurement and are a model for the kind of robust characterization of reproducibility that the scientific community strives for more generally.

Leslie McIntosh (executive director of the U.S. Research Data Alliance and founder and chief executive officer of Ripeta, LLC) discussed potential measures of reproducibility based on natural language processing (NLP) and machine learning (ML). In other words: Can we determine the level of confidence in a research result based on semantic analysis of the documents describing the research, such as publications and the data associated with them? Starting with publications in the health sciences, McIntosh first defined the elements of reproducible research workflows and then sought to automate detection of those elements in the associated articles. The Ripeta Framework[10] uses over 100 variables to characterize research publications in the areas of bibliography, databases and data collection, data mining and cleaning, data analysis, and data sharing and documentation. The system is initialized through manual annotation of a corpus of documents and then trained through NLP and ML methods. Initial results showed that most research papers fall short in terms of transparency to the experimental methods used and are even worse in terms of accessibility of the supporting data. Hopefully, by exposing the deficiencies in the scholarly publication process and the advantages to better sharing of data and data-flow processes, the level of reproducibility will increase.

[9] https://kcdb.bipm.org/
[10] See https://demo.ripeta.com for a demonstration.

Sir Jim Smith (director of science at the Wellcome Trust) presented Wellcome's perspective and approach to reproducibility. It is a topic of great importance to them, and as a funder of research, they expect the people they support to adhere to high standards of research integrity and rigor, including the ways in which research is planned, performed, reported, and shared. This ethos is integral to their Science team strategy "Improving health through the best research," which has four principal aims: create knowledge, strengthen research capability, ensure knowledge is used effectively, and contribute to an environment in which research can flourish.[11] He articulated the meaning of reproducibility in research as

> If you do the same experiment exactly the same way twice, you should get exactly the same result. And if you don't, you've done it differently (perhaps without knowing it), you have made a mistake (perhaps without knowing it), or you have manipulated your data.

He drew attention to a paper "The Economics of Reproducibility in Preclinical Research" by Freedman *et al.*, which estimated that over half of the investment in preclinical research, around $28 billion, in the United States is not reproducible [12]. The causes of this were categorized as biological reagents and reference materials (36 %), study design (28 %), data analysis and reporting (26 %), and laboratory protocols (11 %). Some of the remedies include more detailed reporting of reagents and cell lines, including confirmations of purity. A discussion of the likely contaminants is now becoming mandatory in some journals. This topic was discussed more extensively by Natalie DeSouza (see below). However, some confounding factors are hard to identify *a priori*. An example was cited from Sorge *et al.* [13], who found that exposure of mice and rats to male but not female experimenters produced pain inhibition, and therefore the sex of the experimenter can affect apparent baseline responses in behavioral testing. Smith identified some important factors that can help, including training in experimental design, data handling, and statistics; smaller laboratory sizes (since it can be hard to achieve sufficient oversight in large research groups); preregistration of protocols; training reviewers to spot mistakes; and conducting post-publication reviews to correct mistakes. He raised the topic of data manipulation and commented that figure manipulation is thought to occur in 1 in 40 papers; journals are getting better at spotting it, but provision of raw data is an important safeguard. Training in research integrity is important to ensure that researchers do not use p-hacking or HARKing. Community initiatives such as the San Francisco Declaration on Research Assessment[12] (DORA) are important to rebalance journal hegemony. Encouraging the publication of useful negative results can be a great help to the community, and the new Wellcome Open Research[13] journal facilitates this endeavor. It uses an open-research publishing model with immediate publication followed by open invited peer-review to support reproducibility and transparency.

Natalie De Souza (editor-in-chief, *Nature Methods*) gave a publisher's perspective. Efforts need to be made to reduce irreproducibility due to cherry picking of results, uncontrolled experimenter bias, poor experimental design, statistical problems, overfitting of models to noisy data, faulty reagents, and inappropriate data presentation, amongst others. The causes of irreproducibility are multifaceted, and there is no single remedy. At the local researcher or principal investigator level, this can include training, laboratory management, leadership, and mentoring, while at the research institution and scientific community scale, incentives for rigor and good laboratory leadership are important. She identified four themes that publishers can help with: education and awareness, policy, infrastructure, and improved incentives. For example, Nature Publishing Group (NPG) has been a vocal forum for publishing views and creating debate on reproducibility, and 16 articles were highlighted. An editorial in *Nature Methods* "Better Research through Metrology," based on this workshop, encouraged their readers to consider whether principles of measurement science could have a role to play in their own disciplines [14]. There are also

[11] See https://demo.ripeta.com for a demonstration.
[12] https://sfdora.org
[13] https://wellcomeopenresearch.org

many resources available to help researchers, including *Statistics for Biologists*[14] and *Visual Strategies for Biological Data*.[15] A new digital open-access data journal, *Scientific Data*, provides an infrastructure for reference data sets and standards. The data sets are peer-reviewed and citable, so that authors can get credit for sharing research. Accessible protocols are an important tool to improve reproducibility, and NPG provides *Nature* protocols (peer-reviewed) and a protocol exchange (freely available). Examples of the policies that NPG has introduced include a detailed checklist to accompany papers and elimination of length limits for on-line methods sections. Also, requirements have been introduced to define the sample size of data, to identify whether data are from single or multiple measurements, and to give details of statistical tests and visualization in figures. All *Nature* journals now publish a data availability statement and encourage data citation. In the fourth theme, "shift the incentives," *Nature* journals have mandated author contribution statements to clarify credit and accountability. The use of ORCID is encouraged to provide persistent unique identifiers to researchers. They have also worked to promote article-level metrics rather than journal impact factors (see San Francisco Declaration on Research Assessment, noted previously). In summary, De Souza highlighted that the role of journals is to raise awareness and educate, catalyze and facilitate discussions, and help drive changes. She also stated that journals need to ensure full reporting, effective review, and measured conclusions and also ensure detailed and accurate credit for contributions.

Anne Plant (fellow, Biosystems and Biomaterials Division, Material Measurement Laboratory, NIST) gave the workshop summary. She began by noting that reproducibility of research is *not* a guarantee of accuracy or truth, and that the failure to reproduce a result can be caused by varying but unrecognized experimental conditions. The focus should be less on reproducibility and more on confidence in measurement. Reproducibility issues have led to widespread and sometimes sensational press coverage, but this has negative consequences, such as the erosion of public confidence in science and a lack of scientific, evidence-based decision making. The NMIs can and should play a role in promoting sound design of experiments and robust analysis methods, including proper characterization of uncertainties and sharing of the data underlying scientific conclusions, building on the trust the research community has in the NMIs for their independence and integrity. She identified 16 areas in which the NMIs can help:

- Make it easier to collect protocol details.
- Qualify software.
- Be more engaged with publishers and editors.
- Promote data stewardship and software engineers as a professional position.
- Commit to long-term data preservation.
- Inspire the rest of the scientific community through communications.
- Apply the FAIR principles, and make data shared by the International Committee of Weights and Measures (CIPM) machine readable.
- Provide education and educational materials on SI units and metrology.
- Engage national academies to provide imprimatur.
- Provide trusted algorithms, models, and software.
- Develop workflows and provide best practices and leadership of good stewardship.
- Deploy and support electronic laboratory notebooks and collection of metadata about theoretical/computational experiments.
- Promote the provision of metadata and supporting data at sufficient granularity.
- Develop ways of qualifying (confidence value) and/or understanding the uncertainties of the results of artificial intelligence (AI) and machine learning (ML) analyses.
- Provide domain-specific ground truth data sets.
- Collect all provenance, including operations on data.

[14] https://www.nature.com/collections/qghhqm/pointsofsignificance
[15] https://www.scientificamerican.com/products/nature-products/nature-collections-visual-strategies-for-biological-data/

Her final summary statement was:

> It is important that the public have confidence in the scientific method, and that all researchers, research reviewers, and funders have a good understanding of the hallmarks of scientific investigations that produce results with a high level of confidence.

4. Road-Map Session

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 or endorsed by the individual NMIs.

Intercomparisons and Replication Studies

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

Repeatability and Reproducibility

- Some fields, such as the pharmaceutical industry, require comparisons prior to approval of a new drug.
- Research instruments need to provide readily accessible metadata for all information affecting data and measurements, preferably in open, nonproprietary formats.
- The NMIs should lead by example, demonstrating best measurement practices internally and sharing these with the broader research community.
- Data acquisition should be automated so as to minimize potential for human error.
- Automatic capture of the research process (workflow) should be implemented through to publication of machine-actionable research articles.
- Scientists who produce reproducible, reusable research data and software should be rewarded.

Training

- NMIs should advocate for training in the principles of metrology, uncertainty characterization, statistical methods, and machine learning in university curriculum.
- NMIs could consider assisting in developing best-practice guidelines, providing open-source data sets for training and demonstration of proficiency, and creating a universal platform for access to training materials.
- NMIs should collaborate with data science training programs sponsored by CODATA.
- NMIs should host metrology hackathons for uncertainty estimation in AI and ML.
- NMIs should establish Software Carpentry–like[16] program for exposure to sound measurement methodologies.

International Standards for Data

- Use, adapt, and adopt existing metadata standards.
- Aim for fewer standards, but each with higher adoption rates.
- Assure that standards incorporate proper metrology (*e.g.*, unambiguous expression of units of measure).
- Develop a comprehensive directory of relevant standards and their purpose/scope.

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

Reference Materials and Reference Data

- Rectify the major gap in reference materials in the biomedical and materials science research areas.
- Encourage broader use of reference materials and reference data to 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 ML.

5. Recommendations and Actions

The workshop consensus was that NMIs are uniquely placed to improve reproducibility in research owing to their expertise in measurement and associated measurement uncertainties and their role as impartial and independent bodies. The NMIs also have a responsibility to be role models for the research community and to support their endeavors through leadership of intercomparisons (replicability) studies. Table 1 lists the key recommendations and actions to be taken by the metrology community moving forward.

Table 1. Key recommendations and actions to be taken by the metrology community.

Key Recommendation	Actions
1. The metrology community should seek to reflect and communicate how best practices from the community can contribute to greater confidence in research findings.	• NIST, NPL, and colleagues will prepare a summary of the workshop for publication (this publication). • The NMIs have an important role to lead intercomparisons and proficiency exercises; and to encourage participation from industry and academia. More could be done if funding were available for such activities. • Metrology institutions should be more permeable and open to interaction with industry and academia. Helping to ensure better measurement in research could resolve genuinely different results that lead to new insight from a background of repeatability or reproducibility issues. • The NMIs should have a greater role in doctoral training programs, which help to instill metrology in the next generation of scientists. This also provides an opportunity to take a multidisciplinary and collaborative approach. • Provide reference data sets for algorithm testing.
2. The NMIs should be a role model for the Findable, Accessible, Interoperable, and Re-usable (FAIR) principles, providing public access to data and methodologies to record data provenance.	• Recommend that the CIPM establish a crosscutting advisory committee to address metrology issues arising from increasing volumes of data and specifically to consider how the FAIR principles could be embedded in the activities of the International Bureau of Weights and Measures (BIPM), regional metrology organizations (RMOs), and the wider international metrology community.
3. Due to the rapid development of digital manufacturing (Industry 4.0, *etc*.) and AI, machine-readable methods and protocols, as well as the transfer of digital calibration certificates, should become standard practice.	• Provide the internationally accepted and standardized infrastructure for provenance of data, digital calibration certificates, and accepted ontologies for machine-readable methods. This will require cross-disciplinary and cross-sectoral efforts, which may be embedded in the CIPM advisory committee mentioned above.
4. Professional development of data scientists in a non-classical research role (e.g. data steward, data analyst, data engineer) should be supported.	• Develop plans for career paths that are not covered by the traditional research track. This is already an important topic in the life sciences, and there are opportunities to work together. • NMIs have a long-term role (*e.g.*, SI system) and should engage with data organizations to find solutions for long-term sustainability.
5. The equivalent of “Google Scholar” for data reuse statistics based on digital object identifiers (DOIs) should be created; such a system that credits data authors through citation would drive behavior to better sharing of data.	• Several organizations are exploring ways to improve credit for data sharing, and the metrology community should continue to work with them. We can increase awareness within the NMIs and develop criteria for promotion accordingly.

A summary of the workshop and the recommendations above were presented to the CIPM at its meeting in June 2018. The contribution of the workshop was welcomed, and, in response, CIPM agreed to establish an *ad hoc* working group to consider the role of metrology in improving the reproducibility of research data and related topics. Following the election of the new CIPM at the General Conference of the Metre Convention in November 2018, this working group has been formed with the task of considering the role of metrology in improving the reproducibility of research data and the broader issue of metrology in the wider digital economy.

6. Glossary

AI	Artificial intelligence
BIPM	Bureau International des Poids et Mesures (International Bureau of Weights and Measures)
CIPM	International Committee for Weights and Measures
CSIRO	Commonwealth Scientific and Industrial Organization (Australia)
CODATA	Committee on Data for Science and Technology, under ISC
EDI	Electronic data interchange
FAIR	Findable, Accessible, Interoperable, Reusable
HARKing	Hypothesizing after the results are known
ISC	International Science Council
KRISS	Korea Research Institute of Standards and Science
ML	Machine learning
NIBSC	National Institute for Biological Standards and Control (U.K.)
NLP	Natural language processing
NMI	National metrology institute (e.g., NPL, NIST, PTB, *etc.*)
NSF	National Science Foundation (USA)
PTB	Physikalisch-Technische Bundesanstalt (Germany)
RMO	Regional metrology organizations
SI	Système International (International System of Units)
SRD	Standard reference data
SRM	Standard reference material
SSN	Semantic Sensor Network
STFC	Science and Technology Facilities Council (U.K.)
UCUM	Unified Code for Units of Measure
UKRI	U.K. Research and Innovation
VAMAS	Versailles Project on Advanced Materials and Standards
WC3	World Wide Web Consortium

Acknowledgments

The authors are grateful to the invited speakers listed in the text for making their presentations available and to the other members of the workshop organizing committee: Martyn Sené, NPL (U.K.), J. T. Janssen, NPL (U.K.), Michael Oldham, NPL (U.K.), Alistair Forbes, NPL (U.K.), Stephen Ellison, LGC Group (U.K.), Hendrik Emons, European Commission (Belgium), Saascha Eichstaedt, PTB (Germany), Christian Schneider, National Institute for Biological Standards and Control (U.K.), and Hyun Kyoon Lim, KRISS (South Korea).

7. References

[1] Nosek BA, Aarts AA, Anderson CJ, Anderson, JE, Kappes, HB (2015) Estimating the reproducibility of psychological science. Science 349:943. http://doi.org/10.1126/science.aac4716

[2] Harris R (2018) Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions (Basic Books, New York, NY). ISBN-13 9781541644144.

[3] National Academies of Sciences, Engineering, and Medicine (2019) Reproducibility and Replicability in Science (The National Academies Press, Washington, D.C.). https://doi.org/10.17226/25303

[4] Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454. https://doi.org/10.1038/533452a

[5] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis JPA (2017) A manifesto for reproducible science. Nature Human Behavior 1:0021. https://doi.org/10.1038/s41562-016-0021

[6] Plant AL, Locascio LE, May WE, Gallagher PD (2014) Improved reproducibility by assuring confidence in measurements in biomedical research. Nature Methods 11:895–898. https://doi.org/10.1038/nmeth.3076

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

[8] Sené M, Gilmore I, Janssen JT (2017) Metrology is key to reproducing results. Nature 547:397–399. https://doi.org/10.1038/547397a

[9] Boulton G, Hodson S (2017) CODATA Prospectus: Strategy and Achievement, 2015–2017 (CODATA, Paris, France), 3rd update. https://doi.org/10.5281/zenodo.1167846

[10] Hey T, Tansley S, Tolle K (2009) The Fourth Paradigm: Data-Intensive Scientific Discovery (Microsoft Research, Redmond, WA). ISBN 9780982544204.

[11] Bureau International des Poids et Mesures (BIPM) (2012) International Vocabulary of Metrology (VIM) (BIPM, Sèvres, France), 3rd Ed. https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf

[12] Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. PLOS Biology 13(6):e1002165. https://doi.org/10.1371/journal.pbio.1002165

[13] Sorge RE, Martin LJ, Isbester KA, Sotocinal SG, Rosen S, Tuttle AH, Wieskopf JS, Acland EL, Dokova A, Kadoura B, Leger P, Mapplebeck JC, McPhail M, Delaney A, Wigerblad G, Schumann AP, Quinn T, Frasnelli J, Svensson CI, Sternberg WF, Mogil JS (2014) Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nature Methods 11:629–632. https://doi.org/10.1038/nmeth.2935

[14] De Souza N (2018) Better research through metrology. Nature Methods 15:395. https://doi.org/10.1038/s41592-018-0035-x

About the authors: *Robert J. Hanisch is the director of the Office of Data and Informatics, Material Measurement Laboratory, at NIST. Ian S. Gilmore is a senior NPL fellow and head of science at the National Physical Laboratory, Teddington, United Kingdom. Anne L. Plant is a fellow in the Biosystems and Biomaterials Division, Material Measurement Laboratory, at NIST. The National Institute of Standards and Technology is an agency of the U.S. Department of Commerce.*