ANNE L. PLANT

Build Confidence in Science by Embracing Uncertainty Rather Than Chasing Reproducibility

At a meeting of the American Society of Cell Biology in 2012, I sat in a packed meeting room. The speaker was Glenn Begley, author of a new article reporting that the results from dozens of academic research papers had failed to reproduce after considerable efforts in his company's labs. And these weren't just any research results—they were apparent breakthroughs published in prominent research journals.

The feeling in the room was that this was evidence of an emergency of epic proportions, a crisis of irreproducibility in science. Discussion turned to the idea that perhaps a third party should be required to check the reproducibility of all studies before the results get published. The line to ask questions was long, so I kept my seat. Finally, a colleague for whom I had great respect as a leader in the field asked the rhetorical question that was on my mind: "And who checks the checkers?" In other words, "reproducible" is not the same as "correct." Maybe the "reproducers" would get it wrong. Or maybe both studies were flawed. Or maybe there were enough uncontrolled variables that the two studies were actually performing different experiments.

The assumption that is often made when talking about reproducing scientific results is that one study is right and one is wrong. However, it is likely that (in the absence of fraud) all studies provide useful, although incomplete, information. Within the scientific community, this misleading assumption about reproducibility has led to counterproductive incriminations and wasted resources. And outside the scientific community, blanket concerns about irreproducibility have enabled politicians, activists, and lobbyists to dismiss results they find inconvenient.

Mapping uncertainty

Although a lot of attention has been paid to reproducibility as a way to evaluate the quality of scientific results, a more productive approach would be to assess sources of uncertainty. The primary goal of basic research is to advance scientific knowledge by presenting work so that others can build on it. For many researchers, particularly those doing basic research, reproducing another lab's results is not an efficient way to achieve scientific advances.

This is evident in some of the heroic efforts by laboratories to reproduce other laboratories' results. A recent effort from 50 labs in Brazil confirms significant interlaboratory variability on par with previous reports. Other well-meaning but insufficient activities to combat the "reproducibility crisis" include journal-imposed checklists and research institute guidelines. These efforts attempt to minimize disparate research results by enumerating a list of required specified experimental variables. But even the best-intentioned list may be incomplete or irrelevant when a particular study contains different or additional variables.

A more systematic and conceptual approach would ensure that researchers consider each study with adequate deliberation about which sources of uncertainty could be relevant, rather than simply conforming to a checklist. By carefully assessing sources of uncertainty, researchers can more easily compare disparate study results and know what uncertainties to address in subsequent studies.

Metrology, the science of measurement, offers a systematic approach to determining uncertainty in a result. Formal metrology defines a measurement as a value plus the uncertainty around that value. Over a hundred years ago, international efforts began to create a formal system of metrology at the Bureau International des Poids et Mesures (BIPM). The BIPM has collectively established the units for electricity, time, and other fundamental physical principles and constants; more recently, it has taken on biological measurements and data science. Its *Guide to the Expression of Uncertainty in Measurement* lays out how to consider uncertainty from factors including bias, statistical methods, physical qualities, and complex experiments with many parameters where uncertainties compound. Formal metrology is a proven approach to achieving confidence in measurement, providing clarity and consensus definitions.

Building on this history, and to make the principles of metrology more digestible for researchers, my colleagues and I at the National Institute of Standards and Technology have proposed a framework for identifying the many potential sources of uncertainty in basic research studies. We consider all aspects of a study: the assumptions that underpin hypotheses, the conclusions drawn, and all the experimental details in between.

Understanding uncertainty is an invaluable tool for discovering why lab results seem to conflict, but trying to achieve reproducibility is rarely an easy process. As a metrologist who has led international efforts in quantifying biological systems, I have witnessed the time and commitment required to achieve comparable results in different labs.

One example is the work of an international group of five government laboratories that were quantifying cellular toxicity from nanoparticles. Initially, each lab observed very different dose response curves to the same nanoparticles. Over years of painstaking work, they were able to identify which aspects of the study were different among their different laboratories. By creating a series of control experiments, they figured out why the results deviated and how to mitigate the variability. Ultimately, each lab was able to demonstrate a similar response curve across a series of increasing concentrations of control nanomaterials, leading the researchers to feel confident that their measurements were comparable to one another's, and therefore meaningful.

This study is a consummate example of how to identify sources of uncertainty. First, the assumptions: The study employed a particular kind of nanoparticle and a particular assay as the foundation for testing nanoparticle toxicity. Assumptions that the selected nanoparticle was toxic, that the assay was relevant, and that the positive control toxic substance was appropriate could have been implicit. However, the authors didn't leave the reader wondering if these assumptions were indeed valid; they cited detailed evidence for why these selections were the most likely to provide specific and unambiguous results.

Some of the most obvious sources of uncertainty are associated with the measurements themselves. For the nanoparticle study, the researchers developed a cause-and-effect diagram (Figure 1) to systematically organize the various sources of experimental uncertainty so that these sources could be considered and mitigated. They then tested the variables in a series of control experiments within a multiwell plate layout, in which more of the wells contained controls than sample replicates. These controls enabled researchers to identify the effects of the most important variables—the concentration of a control toxin, differences in cell numbers, and the relationship between dose of nanoparticles and the optical output.

This study is an extreme example of laboratory harmonization, and while this level of reproducibility is necessary when international agreements are involved, it is not practical or useful for most basic research efforts. However, the principles are logical to all scientists, easy to understand, and already implemented frequently, if informally. The importance of control measurements and reference samples to provide benchmarks against which measurement results can be compared is the basis on which one can relate a result with another result. More complete reporting of these details can have great impact on how efficiently science can build on itself.

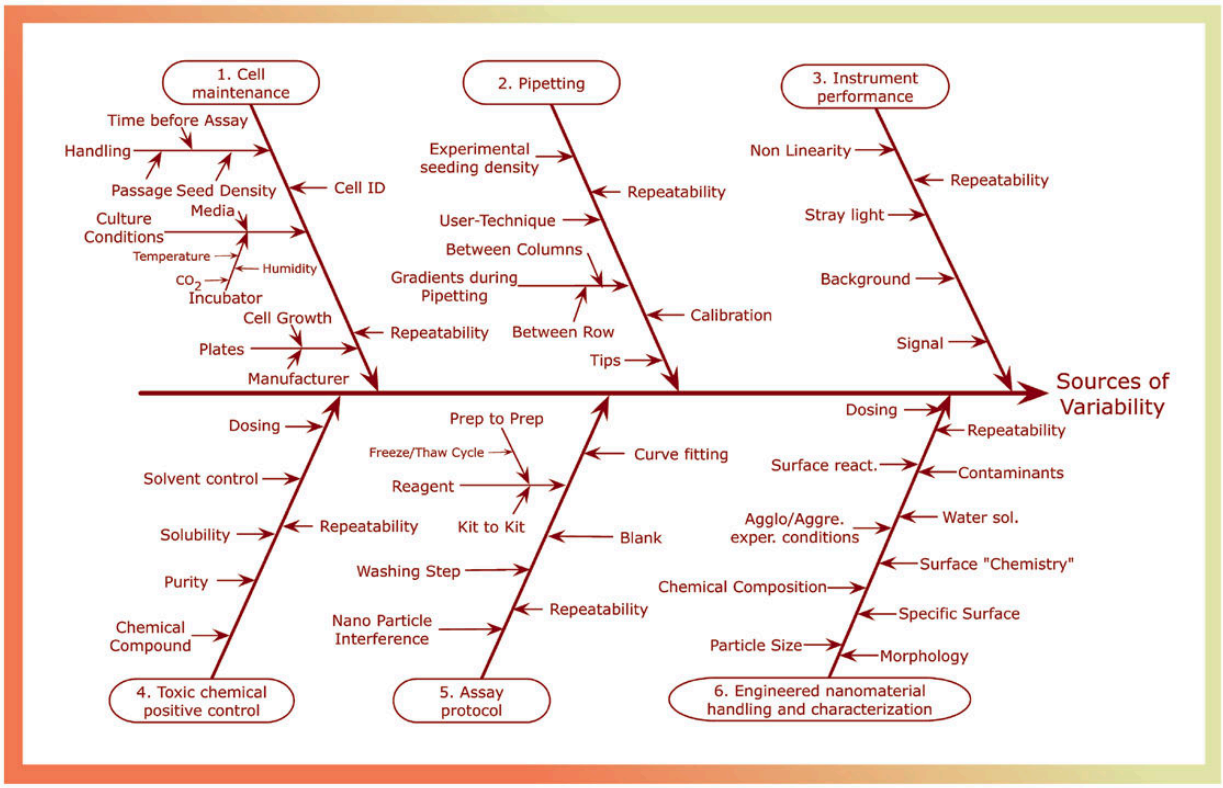
Although these concepts and tools are ingrained in most laboratories, they are not always applied with sufficient intention and are rarely explicitly reported in research publications.

Exploring "variable space"

Of course, it's not feasible to eliminate all sources of uncertainty, but articulating them can bring insight. Exploring what metrologists call *variable space* helps investigators understand how variables influence an observation. This concept can help researchers build on one another's work, but it requires that both intentional and unintentional variables are clearly identified.

As a case in point, a team at *Science* magazine recently hosted a press briefing featuring authors of two papers that appeared to report contradictory results about levels of the amino acid taurine in aging. On closer examination, it became apparent that the studies were examining different age groups and changes over different time

Figure 1.



A cause-and-effect diagram that can help to organize sources of uncertainty in a study so that they can be systematically examined. *Source:* Rösslein, Matthias, et al., "Use of Cause-and-Effect Analysis to Design a High-Quality Nanocytotoxicology Assay," *Chemical Research in Toxicology 28*, no. 1 (2015): 21–30. https://pubs.acs.org/doi/10.1021/tx500327y

ranges. These differences in protocol suggest that the results are not necessarily contradictory—they may in fact be complementary, as additional experiments could make clear.

Identifying variables can also help readers determine how much confidence they should have in a study's results. When methodologies have been refined and validated, assumptions have been verified as robust, statistical sampling is adequate, and the models used have been determined to be appropriate, a study will carry a high level of confidence. Other studies for which there is less confidence may be worthy of reporting, but only if unaddressed sources of uncertainty are made clear.

As more data accumulate from different studies, a scientific consensus may emerge. When new results are presented that are highly divergent from the consensus of previous results, the authors, using principles of metrology, should be able to account for their confidence in the new data and the unexpected conclusions. Scientific conclusions should not be confused with opinions—they must be demonstrably based in tangible criteria that can be compared and validated. This requirement could possibly help counteract the cherry-picking of preferred results by making the requirements for scientific credibility more systematically tangible.

Corralling uncertainty

It has been suggested that more effort and funds should be directed at reproducibility studies. But science would be better served by putting more funding into developing tools that promote data comparability and the reporting of experimental details. Over the years, many efforts to promote data comparability have been pursued by minimally funded, domain-specific, self-organized research communities. In the early 2000s, concerns about what would now be called irreproducibility led to "minimum information about" efforts, such as those for microarray experiments and proteomics experiments.

Some of these ad hoc community activities have demonstrated their potential to significantly impact the ability of a field to achieve insights from disparate data. For example, in the mid-2000s, I got a call from someone I didn't know at the University of Pennsylvania, asking about something I knew almost nothing about: T-cell assays, procedures to assess numbers or activities of this type of immune cell. That call gave me a front-row seat to the development of

MIATA (Minimum Information About T-cell Assays). Researchers from different sectors and organizations had been converging on evaluation of clinical endpoints and assay harmonization. MIATA recommended reporting critical experimental variables—such as the current patient treatment, age, and how a sample was thawed—to enable unambiguous comparison of results. The field of cancer immunotherapy had previously experienced decades of failures. With limited funding, community activities including MIATA's careful efforts helped enable the first cell- and antibody-based cancer immunotherapies. Fast-forward to 2012, when the first pediatric patient was treated with CAR T-cell therapy developed at the University of Pennsylvania. That patient is still healthy and cancer-free today.

The work that led up to MIATA is an exceptional case. Despite the good intentions of volunteers, poorly funded harmonization efforts have a history of petering out over time, and this work is rarely broadly accepted by the research community at large. A reason for this is the prohibitive amount of effort required to collect laboratory protocol details and to create and conform to harmonized terminology for communicating these details. More investment is needed in the tools and activities that would make it easier to codify the protocols, materials, and models used in scientific investigations.

But we can be optimistic that better tools are on the horizon. I feel confident that software employing large language models (LLMs) can make it easier for individual researchers to identify, harmonize, and reuse descriptive metadata terms and reduce reliance on strict schemas developed by committees. But for now, the challenge of accurate and efficient capture of experimental details is still great and will require development of new and clever tools. One bright spot is Cultivarium, a nonprofit, open-source organization that's developing a system for real-time digital capture and query of activities at the lab bench using video and audio together with an LLM. This product was developed with microbiologists in mind; capturing exact manual manipulations and culture conditions for finicky organisms is often crucial.

Other AI software tools that can assess the differences between studies can eventually help identify the variables that lead to conflicting results and further enable our understanding of complex phenomena. More funding for technological advances like these will enable the capture of details that are important sources of uncertainty.

Boosting public confidence

Embracing uncertainty could lead to better science and build a culture that encourages trustworthy reporting. Novelty should not be a primary criterion for scientific manuscript acceptance, because that encourages scientists to claim a novel result even when uncertainty is high. If evaluating uncertainty became an explicit component of peer review and was prioritized by journals over novelty or the "wow" factor, it could change how we communicate and advance scientific findings. If science is to build efficiently on prior studies, how well the authors account for uncertainty in the study should matter as much as the conclusions. This shift could also encourage science reporters to provide the context of uncertainty behind the studies they cover, giving the public a clearer view of the significance of the findings.

> Scientific conclusions should not be confused with opinions—they must be demonstrably based in tangible criteria that can be compared and validated.

At the same time, funders could facilitate the development of software tools designed to efficiently collect experimental details, harmonize metadata terms, and analyze differences between studies. Such tools could focus attention on how to compare and build upon different studies, revealing science as a continuously improving basis of knowledge rather than an accretion of papers that are either right or wrong.

The assessment of uncertainty is a scientific virtue that the research community should publicly and enthusiastically prioritize. A vocal pursuit of these efforts could help the public better understand how science works by incrementally building knowledge based on prior work. If the scientific community can be better at identifying, reporting, and acknowledging sources of uncertainty in their studies, and can more carefully explain what is well-known versus what is still under investigation, it would speed scientific progress, boost public confidence in science, and reduce the promulgation of specious results.

Anne L. Plant *is an emerita fellow at the National Institute of Standards and Technology.*