Statistical Science
2023, Vol. 38, No. 4, 655–671
https://doi.org/10.1214/23-STS899

Tracking Truth Through Measurement and the Spyglass of Statistics

Antonio Possolo

Abstract. The measurement of a quantity is reproducible when mutually independent, multiple measurements made of it yield mutually consistent measurement results, that is, when the measured values, after due allowance for their associated uncertainties, do not differ significantly from one another. Interlaboratory comparisons organized deliberately for the purpose, and meta-analyses that are structured so as to be fit for the same purpose, are procedures of choice to ascertain measurement reproducibility.

The realistic evaluation of measurement uncertainty is a key preliminary to the assessment of reproducibility because lack of reproducibility manifests itself as dispersion or variability of measured values in excess of what their associated uncertainties suggest that they should exhibit. For this reason, we review the distinctive traits of measurement in the physical sciences and technologies, including medicine, and discuss the meaning and expression of measurement uncertainty.

This contribution illustrates the application of statistical models and methods to quantify measurement uncertainty and to assess reproducibility in four concrete, real-life examples, in the process revealing that lack of reproducibility can be a consequence of one or more of the following: intrinsic differences between laboratories making measurements; choice of statistical model and of procedure for data reduction or of causes yet to be identified.

Despite the instances of lack of reproducibility that we review, and many others like them, the outlook is optimistic. First, because "lack of reproducibility is not necessarily bad news; it may herald new discoveries and signal scientific progress" (*Nat. Phys.* **16** (2020) 117–119). Second, and as the example about the measurement of the Newtonian constant of gravitation, G, illustrates, when faced with a reproducibility crisis the scientific community often engages in cooperative efforts to understand the root causes of the lack of reproducibility, leading to advances in scientific knowledge.

Key words and phrases: Avandia, common mean, fixed effect, COVID-19, Newtonian constant of gravitation, Rosiglitazone, dark uncertainty, heterogeneity, interlaboratory study, meta-analysis, random effects, repeatability, replicability, reproducibility, reproduction number, W boson.

1. INTRODUCTION

This contribution reviews how organized comparisons (interlaboratory studies), and meta-analyses of measurement results obtained in different studies or experiments, and the evaluation of measurement uncertainty that underlies them, can contribute to gauge and improve reproducibility in the physical sciences and in medicine. The nature and role of measurement in the social and behavioral sciences, including the education sciences, and attendant issues of reproducibility, lie beyond the scope of this review.

The use of statistical models and of methods of statistical data analysis are illustrated in several examples involving uncertainty evaluations and the intercomparison of measurement results, highlighting the characterization of reproducibility and indicating the role that the evaluation of measurement uncertainty plays in the process.

The article is intended for statisticians concerned with the assessment of reproducibility in measurement as prac-

Antonio Possolo is NIST Fellow and Chief Statistician, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, U.S.A. (e-mail: antonio.possolo@nist.gov).

ticed in national metrology institutes like the National Institute of Standards and Technology (NIST) of the U.S., as well as in many other laboratories where measurements are made that support the practice of medicine, engineering, environmental studies, forensic investigations and that ensure the quality of food, therapies and industrial products.

The article is also intended for physical scientists, medical doctors, engineers, laboratory technicians and others who make measurements and employ statistical methods to assess reproducibility via interlaboratory studies and meta-analyses, and who also wish to gain some appreciation for how the evaluation of measurement uncertainty underlies the assessment of reproducibility.

Section 2 uses the Newtonian constant of gravitation as an example to explain the meaning of notational conventions that are widely used in metrology but that statisticians may be unfamiliar with, and which are used throughout this contribution.

Since measurement plays a key role in science and technology, both the credibility of scientific results and the reliability of technologies hinge on measurement quality, which is the topic of Section 3.

Section 4 discusses the meaning of "reproducibility" and of related concepts. Section 5 presents a reanalysis, employing contemporary techniques, of a historical data set that John Mandel used to illustrate his pioneering approach to characterize measurement reproducibility and repeatability.

Sections 6 (assessment of the risks of a particular therapy), 7 (estimation of the reproduction number of COVID-19) and 8 (measurement of the Newtonian constant of gravitation) provide additional illustrations of how the statistical intercomparison of measurement results contributes to the assessment of reproducibility.

Section 9 gathers some lessons learned about how the application of statistical models and methods can quantify the reproducibility of the conclusions of scientific studies, and in the process increase their trustworthiness, thereby advancing scientific knowledge.

The title chosen for this contribution alludes to the tracking theory of knowledge developed by Nozick [64] and by Roush [79], at the same time as it evokes the dynamic nature of exploratory and confirmatory statistical data analysis, as they "track" the scent of truth in empirical data, thus fulfilling the allegorical role of a spyglass that delivers reliable knowledge built upon reproducible findings.

2. NOTATIONAL CONVENTIONS

The term *standard uncertainty*, and the notation used to denote it, occur repeatedly throughout this contribution, as in $u(G) = 0.000122\ \mathrm{m^3kg^{-1}s^{-2}}$, which Schlamminger et al. [84] reported as the standard uncertainty associated with a measurement of the Newtonian constant of gravitation made at the University of Zürich, Switzerland, $G = (6.674252 \pm 0.000122)\ \mathrm{m^3kg^{-1}s^{-2}}$ (this is the result labeled UZur-06 in Figure 8). There are three different conventions in play here:

- Since the true value of G is unknown, G is modeled as a random variable whose probability distribution characterizes the uncertainty surrounding its true value, yet without impugning the fact that, according to current understanding, G has had a unique, essentially invariant true value throughout most of the history of the universe [59, 25].
- The standard uncertainty, $u(G)$, is the standard deviation of G's distribution. However, since this distribution also comprises uncertainty contributions that are not expressed in the data, for example, uncertainty in the calibration of measuring instruments; metrology uses a term conceived to be more inclusive than "standard error."
- The expression for the value of G includes the parenthetic notation "6.674252(122)," which is shorthand for "6.674252 ± 0.000122," indicating that the digits between parentheses express the standard uncertainty and affect the same number of trailing digits of the value of G while disregarding the location of the decimal point. This parenthetic notation is commonly employed to report measurement results concisely in the scientific literature, as well as in Sections 4, 7 and 8 of this contribution.

3. MEASUREMENT AND MEASUREMENT QUALITY

3.1 Measurement

Measurement, the same as science generally, aims "*to find out* something" ([34], p. 287) based on empirical evidence and employing methods that peer-review determines to be sound and enable empirically verifiable predictions, to obtain this evidence and to analyze it, yielding results that can be essentially reproduced by others.

In practice, our measured values are approximations to the true values of the properties that we intend to measure. These estimates, alone, are of little value because they provide no assurances about their quality. For this reason, a bona fide measurement result comprises both a measured value and an evaluation of measurement uncertainty.

Broadly conceived, measurement is an experimental or computational process that produces an estimate of the true value of a property of a material or virtual object or collection of objects, or of a process, event or series of events, and satisfies these requirements [93, 70]:

(a) The estimate (measured value) is based on a comparison of the property of interest with a property of the same kind realized in a standard that is recognized as a common reference by the community of producers and users of the measurement result;

(b) The measured value is qualified with an evaluation of measurement uncertainty;
(c) The measurement result (measured value together with its associated measurement uncertainty) is used to inform an action or decision.

As example of the comparison mentioned in (a), consider the Eiffel Tower: saying that it is 330 m tall means that its height is 330 times the length of the meter, which is the unit of length in the International System of Units (SI) [6].

The property that is measured (measurand) can be qualitative or quantitative. The species of the plant in NIST Standard Reference Material (SRM) 3246, *Ginkgo biloba* (Leaves), is a qualitative property. The mass fraction of tin in NIST SRM 3161a, Tin Standard Solution (Lot No. 140917), is a quantitative property whose certified value is 10.011 mg/g.

To satisfy requirement (b), the aforementioned estimate of the mass fraction of tin is qualified with an expression of measurement uncertainty, in the form of a confidence interval ranging from 9.986 mg/g to 10.036 mg/g.

Requirement (c) is exemplified by the decision to accept or reject a shipment of boxes of breakfast cereal, which depends on a measurement result for the mass of cereal in the boxes. This can be the average mass of cereal per box, for example, qualified with an evaluation of the associated measurement uncertainty.

3.2 Measurement Quality

Measurement quality is its trustworthiness: the extent to which measured values approximate the corresponding true values sufficiently closely for the purpose they are intended to serve, and do so with assuredly high confidence [71].

Such trustworthiness requires that measurement results be metrologically traceable to appropriate, widely recognized standards of reference, and that the associated uncertainty be small enough to warrant using the measured value in practice as a proxy for the corresponding true value.

Traceability is a property of a measurement result consisting of a documented series of comparisons that relate the measured value to a standard of reference, with each comparison being qualified by an evaluation of the associated measurement uncertainty [73]. Traceability thus guarantees that 1 kg of coffee weighed and sold in a supermarket in Cali, Colombia, has the same mass as 1 kg of coffee bought in Coimbra, Portugal, up to their respective, associated uncertainties.

Measurement uncertainty is the doubt about the true value of the measurand that remains after making a measurement ([75], p. 14). Bell [5] points out that to characterize the margin of this doubt, we need to answer two questions: “How big is the margin?” and “How bad is the doubt?” For NIST SRM 3161a, the size of the margin is gauged by half the length of the confidence interval aforementioned, and the severity of the doubt is expressed by the probability (5% in this case) that said interval does not include the true value of that mass fraction.

Confidence in measurement results can be strengthened by introducing known measurands in the measurement workflow that are indistinguishable from the materials or products that are being measured. Such *check standards* ([63], 2.1.2) were first used in mass measurement [69]. In general, they can be reference materials or calibrated devices delivering certified values whose associated uncertainty has been evaluated reliably.

The convergence toward a particular value as the same measurand is measured repeatedly over time, in independent experiments, is another indication that knowledge about it is solidifying. The history of the measurements of the speed of light and of the Planck constant are notable examples of such convergence [52, 61].

Confidence in a measurement result is bolstered appreciably if one or several so-called *primary methods* of measurement are employed, and they produce measurement results that are essentially in agreement with one another. A primary measurement procedure is such that it does not require calibration with a reference that delivers the same property that one intends to measure. Digital polymerase chain reaction (dPCR) is a primary measurement method for viral loads in samples of bodily fluids [85], and for many other measurements in molecular biology [67].

Coulometry ([35], Section 17-3) can be a primary method for determining the amount of a substance in a solution, which involves counting the number of electrons consumed in a chemical reaction involving that substance. This measurement method involves reference to standards of time and electrical current, but not to standards for the concentration of the substance [4], 2.9.5.

In summary, measurement provides estimates of values of properties of interest to science and technology using recognized standards as references. Both measurement uncertainty and traceability, which characterize measurement’s reliability and validity, are attributes of measurement quality. The demonstration of mutual consistency between measurement results for the same measurand obtained independently of one another, that is, reproducibility (which we turn to next), is another quality attribute of measurement that bolsters the trustworthiness of measurement results.

4. REPRODUCIBILITY

A search for articles listed in the Web of Science that were published between January 1, 2020, and January 31, 2023, and that include the word “reproducibility” in their titles yielded 2524 results (retrieved on February 2, 2023).

These articles are from a very wide range of fields of science and technology, with the largest numbers relating to medical ethics and brain imaging, which together account for almost 14% of the total.

The epistemic value of reproducibility has long been recognized. Referring to measurement standards, Herschel [36] suggested that they ought to possess the "qualities of invariability, indestructibility and identical reproducibility," as well as "some obvious claim to general acceptation as of common interest to all mankind."

Viewing the issue from a different angle, Munafò et al. [60] argue that the debate around reproducibility, rather than a crisis, is an opportunity "to reflect on which aspects of relevant working practices continue to be effective, which can be improved, and which new ways of working can beneficially be introduced." Similarly, Milton and Possolo [52] point out that "lack of reproducibility is not necessarily bad news; it may herald new discoveries and signal scientific progress."

For example, the CDF Collaboration's reanalysis of observations made at Fermilab's Tevatron collider yielded 80 433(9) MeV/c^2 [19] as an estimate of the mass of the W boson, while the corresponding, previous result based on observations made at CERN's Large Hadron Collider had yielded 80 370(19) MeV/c^2 [18]: their standardized difference is 3σ, which suggests a significant difference.

However, an even more dramatic difference is obtained when the latest measurement result obtained by the CDF Collaboration is compared against the prediction that the Standard Model of particle physics makes for the mass of the W boson, 80 357(6) MeV/c^2 [32]: once standardized, this difference amounts to 7σ, and *Science* declared it to be "an upset to the standard model" [15].

4.1 Terminology

The Oxford English Dictionary defines *reproducibility* as "the extent to which consistent results are obtained when an experiment is repeated." The meaning of "repeated," or the sense in which repetition suffices to warrant reproducibility, requires clarification because it can have different flavors, and also because it encompasses a very wide spectrum of modalities.

Concerning its flavors: "repeating" can mean obtaining the same results again and again, or it can mean obtaining essentially the same results, even if not necessarily exactly the same results, where "essentially" means that the results from different repetitions cannot be distinguished once their respective uncertainties are taken into account.

This jigs with the understanding of *replicability* expressed by Fineberg et al. ([30], p. 3): "Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study."

For example, the DELPHI Collaboration et al. [27] determined the mass of the W boson as 80 336(67) MeV/c^2, while the corresponding determination made by the L3 Collaboration [20] was 80 270(55) MeV/c^2. These measurement results are not identical but their difference is not significantly different, either statistically (their standardized difference is 0.8σ) or substantively.

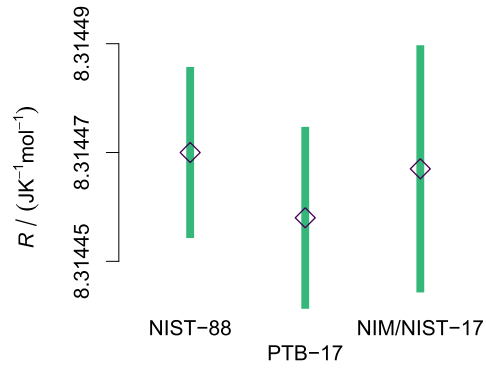

FIG. 1. *Mutually consistent set of three measurement results for the universal gas constant, R, obtained using different measurement methods, from the National Institute of Standards and Technology (NIST) of the United States* (NIST-88), *the Physikalisch-Technische Bundesanstalt of Germany* (PTB-17) *and jointly by the National Institute of Metrology of China and NIST* (NIST/NMI-17). *The diamonds indicate the measured values, and each vertical line segment represents a measured value plus or minus the reported standard uncertainty* (1σ).

Concerning the spectrum of modalities: at one end, we have repetition of the same experiment involving the same materials, apparatuses, methods and procedures, experimenters and place of execution; at the other end, the intended repetition is not of the experiment itself, but of reaching essentially the same conclusions that the original experiment had reached.

This second option in the spectrum of modalities involves measuring the same property, or more generally studying the same phenomenon, but using altogether different approaches, methods and procedures, applied by different experimenters working independently of the original ones, in different laboratories. It is generally agreed that this form of replication has greater epistemic value than the former, because it widens the realm of conditions under which essentially the same conclusions are reached.

For example, Figure 1 shows three measurement results for the universal gas constant, $R = kN_A$, obtained independently of one another and using different measurement methods: k denotes the Boltzmann constant and N_A denotes the Avogadro constant. Two of these measurements were made shortly before the values of k and N_A were fixed as part of the 2019 redefinition of the international system of units [6]. The result labeled PTB-17 was obtained using a dielectric-constant gas thermometer [31], and NIST/NIM-17 was obtained using a Johnson noise thermometer [76]. The result labeled NIST-88 was obtained much earlier, via acoustic gas thermometry [57].

The meaning of reproducibility varies considerably across the scientific literature. Gundersen ([33], Table 1) mentions no fewer than sixteen published, different definitions of reproducibility, recognizes that there are different types and levels of reproducibility, and proposes this definition: "the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators."

The National Academies of Science, Engineering and Medicine (NASEM) use *reproducibility* as synonymous with *computational reproducibility* ([30], p. 4) and define it as "obtaining consistent results using the same input data, computational steps, methods and code, and conditions of analysis." In this sense, reproducibility is less demanding than *replicability*, which NASEM defines as "obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data."

Plesser [68] emphasizes the terminology prevailing in chemistry and in measurement science, which inspired the understanding of *repeatability*, *reproducibility* and *replicability* originally adopted by the Association for Computing Machinery (ACM).

5. QUANTIFYING REPRODUCIBILITY AND REPEATABILITY

Well before the reproducibility "crisis" became a topic of conversation, for example, in a briefing entitled "Trouble at the lab," which *The Economist* published on October 18, 2013, John Mandel [46], a statistician working at the National Bureau of Standards (which became the National Institute of Standards and Technology in 1988), defined *repeatability* as "the variability (or rather smallness of variability) between replicate results obtained on the same material within a single laboratory," and *reproducibility* as "the (smallness of) variability between results obtained on the same material in different laboratories," adding that "more exact definitions are needed."

We will review Mandel's concept of these "more exact definitions" in a reanalysis of the results of an interlaboratory study employing contemporary models and methods of statistical data analysis. The study produced 364 determinations of the stress at 600% elongation, of $I = 7$ different specimens of natural rubber, obtained by $J = 13$ laboratories, each of which made $K = 4$ replicated determinations for each specimen ([46], Table 1). These determinations, and the R code used to analyze them, are listed in the Supplementary Material for this article [72].

The model we shall adopt for these determinations is a linear, mixed-effects model,

$$y_{ijk} = \mu_i + \lambda_j + \epsilon_{ijk}, \tag{1}$$

where μ_i denotes the true mean value of the stress for material i, the $\{\lambda_j\}$ denote laboratory ("random") effects

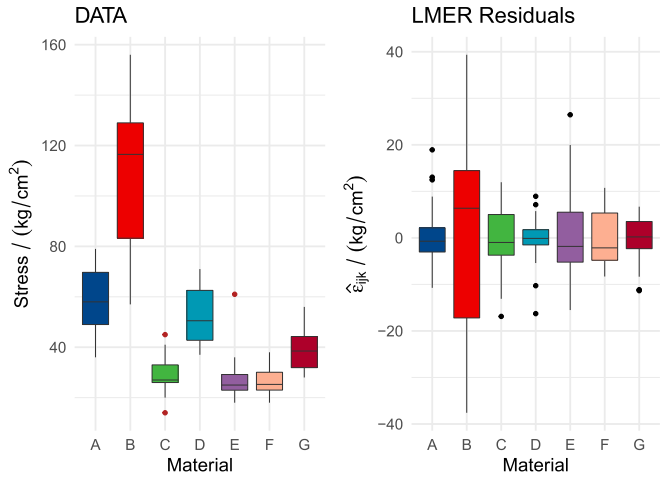

FIG. 2. Left panel: *Boxplots for the raw values of stress at* 600% *elongation of the* 7 *rubber specimens* (A, ..., G), *determined by* 13 *laboratories. Each boxplot summarizes* $13 \times 4 = 52$ *determinations of stress*. Right panel: *Corresponding boxplots of the residuals from the linear, Gaussian mixed effects model fitted to the determinations using R function* `lmer` *defined in package* `lme4` [3].

with mean 0 and standard deviation τ and the $\{\epsilon_{ijk}\}$ denote measurement errors with mean 0 and standard deviation σ, for material $i = 1, \ldots, I$, laboratory $j = 1, \ldots, J$ and replicate $k = 1, \ldots, K$. Owing to the marked heteroscedasticity of the raw values of stress (Figure 2), we will conduct all the analyses using the logarithms of the observed values of stress.

Discussing the presence of apparently outlying observations in interlaboratory studies, Mandel ([47], p. 111), points out that "There is a great temptation to reject such outliers, that is, to discard them from the data prior to calculating precision or accuracy parameters," and adds: "We do not recommend rejection on the basis of purely statistical considerations. Our main reason is that while such rejection procedures always improve the appearance of the data, for example, by drastically reducing the standard deviations, they do nothing in terms of avoiding future instances of outlying results. They have simply sharply reduced the field to which the inferences from the study apply. [...] It is our opinion that the blind application of tests of significance to interlaboratory data for the purpose of rejecting outliers is logically invalid and practically harmful."

We have expressed similar reservations about rejecting measurement results based on "purely statistical considerations" [41, 74]. The Analytical Methods Committee of the Royal Society of Chemistry considered the issue at length more than 30 years ago, and issued recommendations for how not to reject outliers [21, 22].

For the experiment concerned with rubber elongation, in the absence of a substantive reason to reject any of the observations under consideration, we will replace the assumption that the measurement errors $\{\epsilon_{ijk}\}$ are Gaussian,

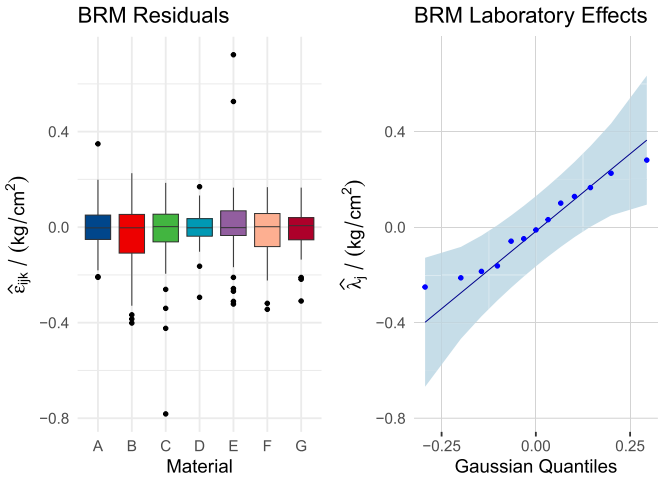

FIG. 3. Left panel: *Boxplots of the residuals from fitting a Bayesian linear mixed-effects model to the logarithms of stress, with Gaussian laboratory effects and Student's t-measurement errors, using R function* `brm` *defined in package* `brms`. Right panel: *QQ-plot of the posterior means for the laboratory effects in the Bayesian mixed effects model with Gaussian random laboratory effects and Student's t-measurement errors.*

with the assumption that they are a sample from a Student's t-distribution whose number of degrees of freedom will be estimated in the course of fitting the model to the data, in the version of the model where the $\{y_{ijk}\}$ in equation (1) denote the logarithms of the observed values of stress.

Figure 3 shows boxplots of the residuals and a QQ-plot for the posterior means of the laboratory effects corresponding to the aforementioned mixed effects model fitted using a Bayesian procedure implemented using R function `brm` defined in package `brms` [13, 14], with the `student` family specification, using Stan [16, 87] and R [86] codes listed in the Supplementary Material [72].

The prior distributions for the (fixed) effects attributable to differences between rubber specimens were essentially noninformative Gaussian distributions. The priors for τ and σ were half-Cauchy distributions. A single σ as standard deviation for all the measurement errors seems justified by the sufficient homoscedasticity apparent in the left panel of Figure 3, and the assumption of Gaussian laboratory effects is justified by the QQ-plot in the right panel of the same figure. The prior distribution for the number of degrees of freedom, ν, of the Student's t-distribution for the $\{\epsilon_{ijk}\}$ was gamma such that with 95% prior probability, $1 < \nu < 45$.

Figure 4 shows posterior probability density estimates of the laboratory effects, indicating that several of the laboratory effects differ significantly from 0, hence that there is significant heterogeneity (between-laboratory variability), or *dark uncertainty* [89], that is, the laboratory averages, once adjusted for the effects of the different rubbers, are more dispersed than they should be considering the

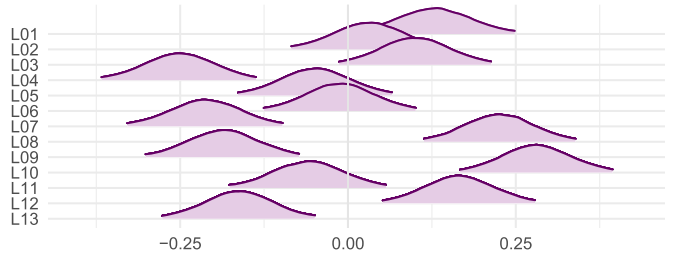

FIG. 4. *Posterior probability density estimates of the laboratory effects, $\{\lambda_j\}$. Each shaded area amounts to 95% of the posterior probability.*

variability of the replicated determinations that individual laboratories made on each rubber.

Figure 5 shows estimates of the posterior densities of the number of degrees of freedom (ν) for the Student's t-distribution of the measurement errors, and also of the between-laboratory (τ) and within-laboratory (σ) standard deviations, which we will use next to quantify the repeatability and reproducibility achieved in this study. The mean of the posterior distribution of the number of degrees of freedom of the Student's t-distribution adopted for the measurement errors, $\{\epsilon_{ijk}\}$, was 2.7(5).

Mandel ([46], p. 78) quantified repeatability in terms of "a quantity that will be exceeded only about 5 percent of the time by the difference, taken in absolute value, of two randomly selected test results obtained in the same laboratory on a given material." Here, *test result* means an average of 4 replicated determinations that a laboratory makes for a rubber specimen. In this conformity, (lack of) *repeatability* is quantified as

$$r = 2\sqrt{2}\hat{\sigma}/\sqrt{K} = 2\sqrt{2}(0.072)/\sqrt{4} = 0.10,$$

and (lack of) *reproducibility* is quantified as

$$R = 2\sqrt{2(\hat{\tau}^2 + \hat{\sigma}^2/K)}$$
$$= 2\sqrt{2(0.185^2 + 0.072^2/4)} = 0.53.$$

In the expressions for both r and R, the first "2" is the rounded value of the 97.5th percentile of the standard Gaussian distribution. $\hat{\sigma}$ and $\hat{\tau}$ denote the medians of

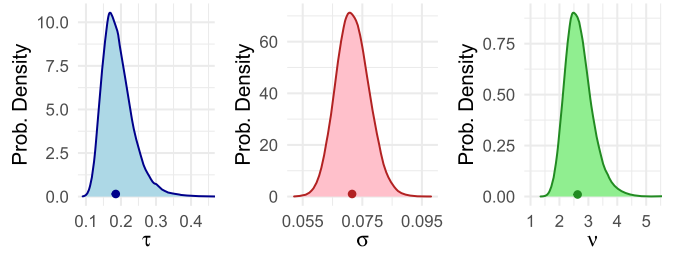

FIG. 5. *Posterior probability density estimates of the between-laboratory (τ) and within-laboratory (σ) standard deviations, and of the number of degrees of freedom (ν) for the Student's t-distribution of the measurement errors. The dots indicate the posterior medians.*

the respective posterior distributions; they are unitless because the analysis is being done using the logarithms of the values of stress, and the logarithm "swallows" units as can be seen by its series expansion presented in [65], 4.6.4.

It is important to realize that these quantifications of repeatability and of reproducibility are supported by different amounts of evidence. In fact, the evaluation of repeatability is based on the variability of 13 groups of 28 individual determinations of stress each (whose logarithms have approximately constant variance), while the evaluation of reproducibility is based on the variability of 91 averages (13 for each of 7 rubber specimens).

Mandel ([46], p. 79) noticed that the different amounts of evidence that support the evaluations of repeatability and reproducibility can be captured using the following fact pointed out by Blackman and Tukey ([8], p. 208): if V is a multiple of a chi-square random variable with m degrees of freedom, for example, when V is an estimate of a variance component, then its coefficient of variation, CV, is $\sqrt{2/m}$. For this reason, Blackman and Tukey [8] propose $2/(\mathrm{CV})^2$ as an equivalent number of degrees of freedom (also called *degrees of firmness* [9], p. 290) supporting V.

To compute the degrees of firmness of the repeatability, r, and of the reproducibility, R, one can either simply compute their respective coefficients of variation based on the MCMC samples drawn from the posterior distributions of σ and τ, or possibly better, employ an analog of the coefficient of variation that may be less sensitive to the asymmetry of these posterior distributions, whose densities are depicted in Figure 5. In this particular case, the two options produce very similar assessments of the degrees of firmness of r and of R.

The robust version of the degree of firmness for r is computed as the ratio between half the length of a 68% credible interval for σ centered at the posterior median of σ, and this posterior median. The value of this ratio is 338. The robust version of the degree of firmness for R, defined similarly, is 48. Hence, and not unexpectedly, the evaluation of repeatability has about 7 times greater firmness than the evaluation of reproducibility.

In general, repeatability depends both on the measurand and on the particular laboratory making the measurements, while reproducibility depends on the measurand and on the class of laboratories that the laboratories participating in the study actually represent.

Also in this case, the logarithmic transformation of the values of stress, together with the adjustment for differences between the rubber specimens accomplished by the mixed effects model, achieved sufficient homoscedasticity within-laboratories, and also enabled using a single τ to quantify the between-laboratories variability, so as to justify pooling the results and producing single evaluations of repeatability and reproducibility.

This reanalysis shows that contemporary tools for statistical modeling and data analysis, which were not available in John Mandel's time, afford great flexibility for accurate modeling. For example, replacing the assumption that measurement errors are Gaussian with the assumption that they follow a Student's t-distribution can be handled easily in the context of a Bayesian model owing to the availability of Markov chain Monte Carlo sampling.

Also, suitably chosen reexpression (which in this case is as simple as taking logarithms) can go a long way toward simplifying the analysis and increasing the adequacy of statistical models to data ([58], Chapter 5). However, the fundamental insights and specific proposals that John Mandel offered 50 years ago, about how to quantify repeatability and reproducibility, withstood the test of time, and continue to be valuable.

6. ROSIGLITAZONE

On July 22, 2007, *The New York Times* reported that Dr. Steven Nissen's "questioning of the safety of the Avandia diabetes medication in late May" had "prompted a federal safety alert and led to a sales decline of about 30 percent for the drug," which had earned GlaxoSmithKline (GSK) $3.2 billion in 2006.

The basis for that questioning was a meta-analysis [62] of 42 clinical studies of the risk of myocardial infarction and death from cardiovascular causes seemingly associated with the use of rosiglitazone, which is the active ingredient of Avandia. The results of each of these studies can be summarized in a 2×2 table, for example, Table 1 for the ADOPT study [91, 39], which was a randomized, double-blind, parallel-group study involving 4351 patients with recently diagnosed type 2 diabetes.

All together, the 42 studies whose results are listed in Nissen and Wolski ([62], Table 3) involved 27 833 patients. The prevalence of myocardial infarction was around 0.6% in both the rosiglitazone and control groups.

In four of these studies, there were no cases of myocardial infarction either in the rosiglitazone group or in the control group. These four were therefore excluded from consideration by those methods of data reduction

TABLE 1

Results of the ADOPT study, where patients were randomized to receive double-blinded rosiglitazone, glyburide or metformin, and were treated for periods of 4 *years median duration, as originally reported by Kahn et al.* [39], *Table* 2, *and transcribed by Nissen and Wolski* [62], *Table* 3

	Myocardial infarction		
	Yes	No	Total
Rosiglitazone Group	27	1429	1456
Control Group	41	2854	2895

TABLE 2
Estimates and lower (LWR) and upper (UPR) endpoints of 95% confidence intervals for the odds ratio (OR) comparing the effects of rosiglitazone and control on myocardial infarction

	OR	LWR	UPR
Peto	1.428	1.031	1.979
Mantel-Haenszel	1.427	1.029	1.978
Weighted Median	1.300	1.001	2.014
DerSimonian-Laird	1.286	0.940	1.759
REML	1.286	0.940	1.759
Bayes	1.280	0.928	1.762

which we have employed for this reanalysis that take estimates of log odds ratios, and their associated uncertainties as inputs: DerSimonian–Laird [28], REML [81] and a Bayesian procedure detailed below. Since neither Peto's [95] nor Mantel-Haenszel's [49] procedures require the calculation of log odds ratios, they used the results from all 42 studies.

Nissen and Wolski [62] chose Peto's method for their data reductions, which was a very reasonable choice considering the findings reported by Bradburn et al. [12]: that, in a comparative evaluation of the performance of 12 methods for pooling rare events (with event rates below 1%), Peto's method was the least biased and most powerful method, and provided the best confidence interval coverage, provided there was no substantial imbalance between treatment and control group sizes within trials, and treatment effects were not exceptionally large, which is generally the case for these trials that involved rosiglitazone.

Table 2 lists the estimates of log odds ratio, and corresponding 95% confidence intervals, resulting from pooling the results from the trials listed in Nissen and Wolski ([62], Table 3) using five different statistical procedures. The methods of Peto, Mantel-Haenszel, DerSimonian-Laird and REML were applied as implemented in R function `rma` of package `metafor` [92]. Figure 6 depicts the log odds ratios for the different studies and the consensus log odds ratio corresponding to Peto's method.

The model used in the Bayesian procedure corresponding to the last line of Table 2 modeled the log odds ratios as outcomes of Gaussian random variables, with the usual large sample approximation for their standard errors ([38], 9.2). The prior distribution for the mean log odds ratio was centered at 0 and had a large standard deviation (5), and the between-study standard deviation, τ, had a half- Cauchy prior distribution with median 0.05. The posterior distribution of τ had median 0.04. A 95% credible interval for τ ranged from 0.002 to 0.3. The model was implemented using R function `brm` defined in package `brms` [13], as detailed in the Supplementary Material [72].

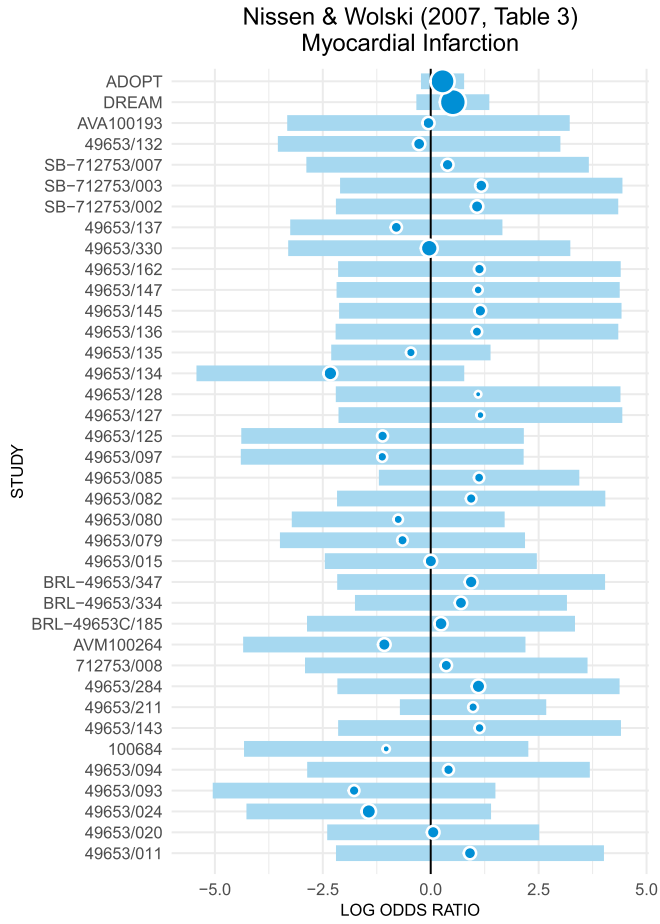

FIG. 6. *Forest plot showing 95% confidence intervals (thick, horizontal (light blue) bars) for the log odds ratios for rosiglitazone versus control in the 38 studies listed in Nissen and Wolski [62], Table 3, that had at least one death in the control group.*

The results for Peto's method (first line in Table 2) reproduce the corresponding results in Nissen and Wolski ([62], Table 4), and the results from the Mantel-Haenszel procedure are in close agreement with Peto's. For neither method does the 95% confidence interval straddle 1. However, the results from the last three procedures—DerSimonian-Laird, REML and Bayes—do not unequivocally corroborate the conclusion of the first two. The *NIST Decision Tree* [74] recommends that the results from the individual studies be combined using the weighted median, which produces the consensus value and 95% confidence interval (based on the non-parametric bootstrap) listed in the third line of Table 2.

The apparently increased risk of cardiovascular events associated with the use of rosiglitazone has been reexamined repeatedly since Nissen and Wolski [62] first rang the alarm bell in 2007, both via critical reanalyses [29] of the same data, and also considering the results of subsequent studies, for example, the RECORD study [37].

Following a recommendation that the European Medicines Agency made on September 23, 2010, to suspend the marketing authorizations for medications containing rosiglitazone, Avandia has been withdrawn from

use throughout the European Union (https://www.ema.europa.eu/en/medicines/human/EPAR/avandia). On July 2, 2012, *The New York Times* reported that "GlaxoSmithKline agreed to plead guilty to criminal charges and pay $3 billion in fines for promoting its best-selling antidepressants for unapproved uses and failing to report safety data about a top diabetes drug" [88]—the diabetes drug was Avandia.

The principal lesson that can be drawn from this example is that different statistical models and methods of data analysis, which may all be comparably adequate for the task at hand, can lead to markedly different conclusions when they are applied to the same data. In this case, three out of the six methods whose results are summarized in Table 2 suggest that the use of rosiglitazone induces a significant risk of myocardial infarction, while the other three do not corroborate such conclusion. Therefore, differences between models and between methods of data reduction can pose a challenge to the reproducibility of research results impacting an issue of the greatest interest in public health.

7. REPRODUCTION NUMBER

The British Health Security Agency (UKHSA) has been publishing consensus values weekly, since May 2020, for the reproduction number, R, of COVID-19. The UKHSA explains it thus: "the reproduction number (R) is the average number of secondary infections produced by a single infected person. An R value of 1 means that on average every person who is infected will infect 1 other person, meaning the total number of infections is stable. If R is 2, on average, each infected person infects 2 more people. If R is 0.5, then on average for each 2 infected people, there will be only 1 new infection. If R is greater than 1, the epidemic is growing, if R is less than 1 the epidemic is shrinking."

The consensus estimate results from blending estimates produced by different research groups, mostly from British universities, working independently of one another and using different models. Blending is done as an exercise in meta-analysis [45].

However, each research group reports several quantiles of the probability distribution that expresses the uncertainty surrounding R, while most procedures used for meta-analysis expect the mean and the standard deviation of R's distribution as inputs. Maishman et al. ([45], Table 1) list the 5th, 25th, 50th, 75th and 95th percentiles for R's distribution, as produced by each of eleven models for a particular (but unspecified) date and region of the UK.

For model 3 in Table 1 of [45], these percentiles are 0.64, 0.70, 0.74, 0.79 and 0.87, respectively. The procedure that Maishman et al. [45] use to derive estimates of the mean and of the standard deviation of R involves consideration of an estimate of the skewness of the distribution based on these percentiles. For these particular percentiles, the procedure reduces to modeling R's distribution as being Gaussian with mean $R = 0.74$ and standard deviation $u(R) = 0.079$.

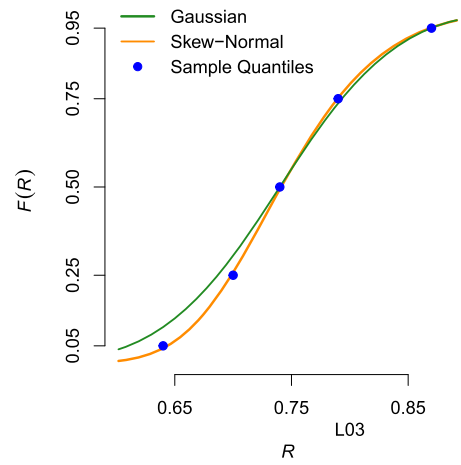

FIG. 7. *Gaussian (green) and skew-normal (orange) approximations to the sample quantiles* $Q(0.05) = 0.64$, $Q(0.25) = 0.70$, $Q(0.50) = 0.74$, $Q(0.75) = 0.79$ *and* $Q(0.95) = 0.87$, *which are represented by the (blue) dots.* $F(R)$ *denotes the probability that the true value of the reproduction number will be less than or equal to* R.

Considering that the eleven sets of percentiles listed in Maishman et al. ([45], Table 1) exhibit fairly mild skewness, we have adopted an alternative modeling approach that approximates the sample percentiles by corresponding percentiles of a skew-normal distribution [1].

The approach involves finding values of the parameters of the skew-normal distribution—ξ (location), ω (scale) and α (shape)—that minimize $\sum_{i=1}^{5}(q_i - \theta_i)^2$, where the $\{q_i\}$ are the aforementioned sample percentiles, the $\{\theta_i = Q(p_i|\xi, \omega, \alpha)\}$ are the corresponding skew-normal percentiles, and Q denotes the quantile function of the skew-normal distribution. Once estimates of ξ, ω and α are in hand, the mean and standard deviation are computed as $\xi + \omega\delta\sqrt{2/\pi}$ and $\omega(1 - 2\delta^2/\pi)$, where $\delta = \alpha/\sqrt{1+\alpha^2}$.

Figure 7 shows the Gaussian cumulative distribution function and its skew-normal counterpart fitted to the percentiles that Maishman et al. ([45], Table 1) list for model 3, showing that the skew-normal model is appreciably more accurate than the Gaussian model. Table 3 lists the means and standard deviations imputed by Maishman et al. [45] for the eleven models, and their counterparts obtained using the skew-normal approximation.

Table 4 reveals details of the differences induced by the two different methods used to impute the mean and standard deviation based on sample percentiles, and also the differences attributable to four different statistical models used to reduce the data to obtain a consensus value and to evaluate the associated uncertainty.

TABLE 3
Estimates and standard uncertainties, $\{R_{G,j}\}$ and $\{u(R_{G,j})\}$, for the values of the reproduction number listed in Maishman et al. [45], Table 1, which are based on a Gaussian model, and their counterparts, $\{R_{SN,j}\}$ and $\{u(R_{SN,j})\}$, based on the skew-normal model, for epidemic models $j = 1, ldots, 12$. Model 8 did not produce results for this reporting period

j	$R_{G,j}$	$u(R_{G,j})$	$R_{SN,j}$	$u(R_{SN,j})$
1	0.74	0.079	0.7435	0.0858
2	0.7045	0.0742	0.7123	0.0388
3	0.74	0.079	0.7466	0.0491
4	0.75	0.2371	0.7576	0.2068
5	0.7954	0.0028	0.7949	0.0020
6	0.8329	0.0256	0.8361	0.0136
7	0.7862	0.1233	0.7895	0.1142
9	0.9382	0.1351	0.9437	0.0899
10	0.8302	0.0077	0.8302	0.0076
11	0.9293	0.0637	0.9314	0.0570
12	0.76	0.0608	0.7572	0.0636

The results listed in Table 4 for the DerSimonian–Laird procedure (DL) [28], the Mandel–Paule procedure (MP) [48] and the restricted maximum likelihood procedure (REML) [81], all were obtained using R function `rma` defined in package `metafor` [92]. The results for the hierarchical Bayesian procedure with Gaussian laboratory effects and Gaussian errors (HGG) were obtained using the *NIST Decision Tree* [74].

TABLE 4
The upper section of the table lists the results of four alternative meta-analyses applied to the means and standard errors imputed using the method described by Maishman et al. [45]. The lower section lists their counterparts for the method that uses the skew-normal distribution. The four procedures (DL, HGG, MP, REML) used to blend the results in Table 3 are referenced in the text. R denotes the consensus estimate of the reproduction number and $u(R)$ denotes the associated standard uncertainty. LWR and UPR are the endpoints of 95% confidence or credible intervals for the true value of R, and τ (dark uncertainty) is an estimate of the standard deviation of the (random) effects attributable to the different models for the epidemic

	R	$u(R)$	LWR	UPR	τ
	Pooling $\{(R_{G,j}, u(R_{G,j}))\}$ from Table 3				
DL	0.8112	0.0135	0.7848	0.8377	0.0229
HGG	0.8092	0.0184	0.7717	0.8467	0.0334
MP	0.8114	0.0125	0.7869	0.8360	0.0206
REML	0.8114	0.0126	0.7868	0.8361	0.0207
	Pooling $\{(R_{SN,j}, u(R_{SN,j}))\}$ from Table 3				
DL	0.8088	0.0137	0.7819	0.8356	0.0269
HGG	0.8072	0.0209	0.7647	0.8497	0.0453
MP	0.8062	0.0194	0.7682	0.8442	0.0441
REML	0.8065	0.0185	0.7702	0.8427	0.0412

Even though none of the differences between the consensus values derived from the $\{(R_{G,j}, u(R_{G,j}))\}$ or from the $\{(R_{SN,j}, u(R_{SN,j}))\}$, using the different blending procedures (DL, HGG, MP, REML), are significantly different from one another; Table 4 does reveal differences worth noting from the viewpoint of reproducibility.

The estimates of the dark uncertainty, τ, in particular, are rather sensitive to the model employed to impute the mean and the standard deviation that correspond to a particular set of percentiles. This is not surprising because it simply expresses the fact that the values of the standard uncertainty, $u(R)$, based on the skew-normal model are generally smaller than their counterparts that are based on the Gaussian model (Table 3).

The estimates of τ also are fairly sensitive to the statistical procedure used for the purpose, for example, HGG's estimate of τ is 1.7 times larger than DL's estimate (first two lines of the lower panel of Table 4). Even though this is not surprising either, considering that τ is a particularly challenging estimand [44, 43], it also influences the evaluation of $u(R)$ [42], thus impacting reproducibility.

The foregoing retrospective of the development of a consensus estimate for the reproduction number of the COVID-19 pandemic reveals that apparently minor differences between fairly simple choices about how to prepare the data for an assessment of reproducibility, can have their effects amplified when different procedures are then used to blend the results in a meta-analysis. In addition, those differences also impact the extent to which the reproducibility of the conclusions depends on the particular procedure employed for the meta-analysis.

8. BIG G

Newton's law of universal gravitation states that two massive objects attract one another with a force that is directly proportional to the product of their masses, and inversely proportional to the square of the distance between their centers of mass: the constant of proportionality is the Newtonian constant of gravitation, G, also informally called "Big G" in contradistinction to "small g," which refers to g, the acceleration of a massive body in free-fall toward the Earth.

G is believed to have the same value everywhere throughout the universe, and figures not only in Newton's third law, but also in the equations of Einstein's theory of general relativity [53]. Big G's lofty status notwithstanding, its relative standard uncertainty, of about 22 parts per million, is much larger than the relative uncertainties of most other fundamental constants [90].

The uncertainty surrounding G is relatively large for three principal reasons: (i) it is not possible to leverage knowledge of the values of other fundamental constants to reduce the uncertainty associated with the estimate of G because there is no known relation between G

and the other fundamental constants; (ii) measuring G is very challenging because it involves measuring extremely small forces and (iii) the measured values of G are appreciably more dispersed than their individual measurement uncertainties intimate.

Reason (iii) is a manifestation of lack of reproducibility, as independent experiments, relying either on different physical principles or on different implementations of the same principle, have historically yielded mutually inconsistent measurement results.

Figure 8 shows the measurement results that CODATA (Committee on Data of the International Science Council) took into account for the 2018 release of the recommended values of the fundamental physical constants [90], and the results of two alternative statistical measurement models and data reductions for them.

Two kinds of statistical models have been used for measurement results such as these, depending on how one addresses their mutual inconsistency. The model discussed in Section 8.1 is based on Birge's [7] suggestion whereby the reported uncertainties are magnified by a factor (*Birge ratio*) sufficiently large to achieve mutual consistency. The model discussed in Section 8.2, which we call the laboratory effects model, is a conventional mixed effects model [50], where G is the fixed effect and the experiment effects are the random effects. Both models will be fitted taking into account the three nonnull correlations between the measured values $\{G_j\}$ listed in the caption of Table XXIX in Tiesinga et al. [90].

Baker and Jackson [2], Koepke et al. [41], Merkatas et al. [51] all compare and discuss these two kinds of models, and point out that the preference for one or for the other seems to be mostly cultural, with CODATA and the Particle Data Group (pdg.lbl.gov) [32] favoring the Birge ratio, while medical meta-analysis [23] and interlaboratory studies in measurement science [80] generally opting for the additive mixed effects model.

The 16 measurement results for G are mutually inconsistent as judged by Cochran's Q test [17], which yields an exceedingly small p-value. Figure 8 also shows the value of G recommended by CODATA in 2018 [90], and the estimates of G obtained by application of the multiplicative and additive models that address such mutual inconsistency, as detailed in the following two subsections.

8.1 Common Mean Model for G

The multiplicative model is a heteroscedastic, Gaussian, common mean model [11] (also called "fixed effect" model—note the singular in "effect," hence a different model from the conventional fixed effects model), which amplifies the standard uncertainties multiplicatively with the inflation factor $\kappa > 0$:

$$G_j = G + \kappa\epsilon_j. \tag{2}$$

The measurement errors $\{\epsilon_j\}$ are assumed to have a joint multivariate Gaussian distribution with mean 0 and the same units as G, whose covariance matrix has the $\{u^2(G_j)\}$ along the main diagonal, and all the off-diagonal entries are 0 except for those that involve the correlations listed in the caption of Table XXIX of Tiesinga et al. [90]: 0.351 between NIST-82 and LANL-97; 0.134 between HUST-05 and HUST-09 and 0.068 between HUST-09 and HUSTT-18.

Both the 2014 [56] and 2018 [90] releases of the values recommended by CODATA for the fundamental constants employ an ad hoc procedure to assign a value to κ, as the smallest positive number such that the resulting, standardized residuals (which Tiesinga et al. [90] call *normalized residuals*) all have absolute values no larger than 2. This choice, which Merkatas et al. ([51], Section 3.2) show is overly conservative, yields 3.9 as estimate of κ.

Both maximum likelihood estimation (MLE) and the Bayesian alternative described by Bodnar and Elster [10] are model-based alternatives preferable to the aforementioned ad hoc procedure to estimate κ.

The maximum likelihood estimates of G and κ in equation (2) are $\widehat{G} = 6.67430(13) \times 10^{-11}$ $\mathrm{m^3kg^{-1}s^{-2}}$ and $\widehat{\kappa} = 3.5(6)$. Note that the maximum likelihood estimate of κ is qualified with an evaluation of the associated uncertainty, which is neither recognized nor propagated for the ad hoc estimate used by Tiesinga et al. [90]. The corresponding results are depicted in the left panel of Figure 8.

8.2 Laboratory Effects Model for G

The *NIST Decision Tree* [74] (which ignores the three correlations aforementioned) recommends a Bayesian hierarchical model with Gaussian random effects and Gaussian measurement errors for these 16 measurement results, similar to the model in equation (1):

$$G_j = G + \lambda_j + \epsilon_j, \tag{3}$$

where the $\{\epsilon_j\}$ are assumed to be independent and Gaussian, all with mean zero and standard deviations equal to the reported standard uncertainties, $\{u(G_j)\}$, all of which are also assumed to be based on very large numbers of degrees of freedom—likely an unrealistic assumption.

The experiment effects, $\{\lambda_j\}$, are assumed to be Gaussian, centered at 0 $\mathrm{m^3kg^{-1}s^{-2}}$ and with a covariance matrix all of whose elements are zero, except for τ^2 along the main diagonal, and the same three elements in the upper and lower triangles that correspond to the three nonnull correlations mentioned above in Section 8.1.

This model is identifiable because the data are the pairs $\{(G_j, u(G_j))\}$: since the $\{\epsilon_j\}$ should be consistent with the $\{u(G_j)\}$, the $\{G_j\}$ being overdispersed relative to the reported uncertainties suggests that the $\{\lambda_j\}$ cannot all be zero.

A Bayesian version of the model in equation (3), taking the aforementioned correlations into account, was fitted to

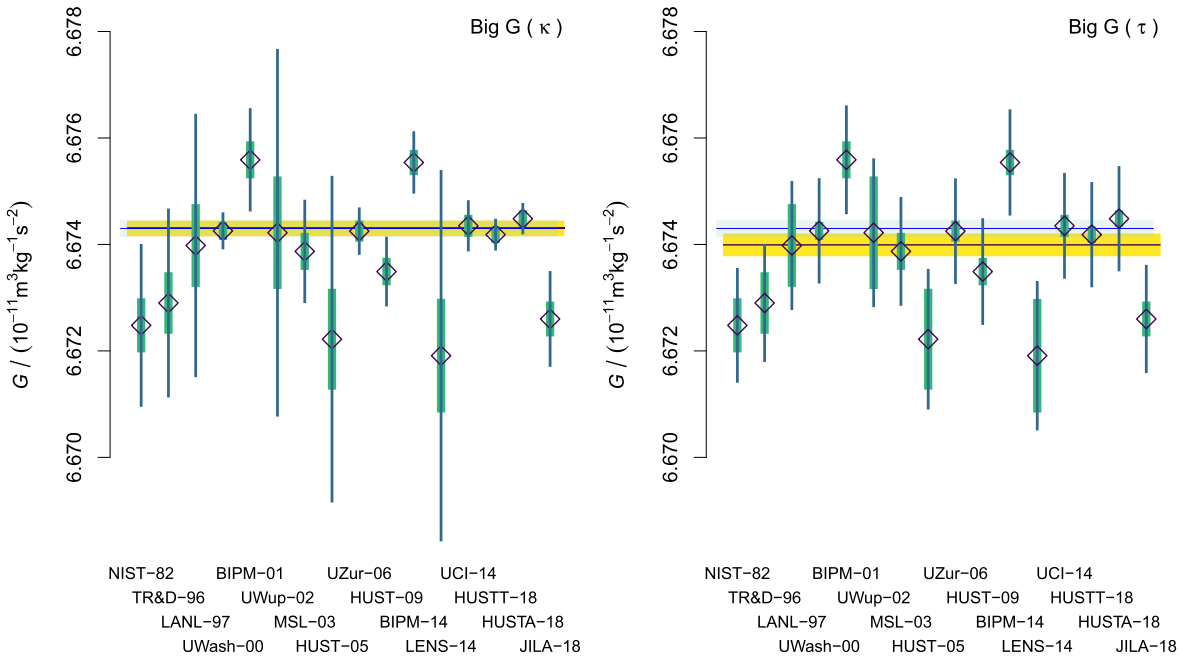

FIG. 8. *Measurement results for G, and results from two alternative statistical models and corresponding data reductions. The labels at the bottom are the same that are used by Tiesinga et al. ([90], Table XXIX), where the corresponding references are listed. The diamonds represent the measured values. The (green) thick vertical line segments represent the measurement results $\{G_j \pm u(G_j)\}$. The (dark blue) thin horizontal line segment, and the light blue band centered on it, represent the* 2018 *CODATA recommended value for G and the associated standard uncertainty* [90], *Section XIX.* Left panel: *The (dark brown) thin horizontal line segment and the yellow band centered on it represent the consensus value computed using the common mean model of equation* (2) *fitted by maximum likelihood, and taking into account the correlations between experiments listed in the caption of Tiesinga et al. ([90], Table XXIX). The (purple) thin vertical line segments represent the $\{G_j \pm \widehat{\kappa} u(G_j)\}$.* Right panel: *Counterpart of the left panel for the mixed effects, Bayesian hierarchical model with Gaussian experiment effects and Gaussian measurement errors, also taking into account the correlations aforementioned. The (purple) thin vertical line segments represent the $\{G_j \pm (\widehat{\tau}^2 + u^2(G_j))^{1/2}\}$ where $\widehat{\tau}$ denotes τ's posterior mean.*

the data listed in Table XXIX of Tiesinga et al. [90] using Stan [16, 87] and R [86] codes listed in the Supplementary Material [72], with the results depicted in the right panel of Figure 8.

The prior distribution chosen for G was Gaussian with mean set equal to the 2014 CODATA recommended value for G [55], and with standard deviation set equal to the corresponding standard uncertainty. The prior distribution chosen for τ was half-Cauchy with median set equal to the MAD (as defined in the R environment for statistical computing and graphics [86]) of the measured values.

The posterior mean of G is $6.67399(20) \times 10^{-11}$ m$^3 \cdot$ kg^{-1}s^{-2}, which is not statistically significantly different from the 2018 CODATA [90] recommended value because the absolute value of their difference amounts to 1.24 times the standard error of their difference. The dark uncertainty, τ, had posterior mean 0.00096×10^{-11} m^{3}kg^{-1}s^{-2}, which is 3.8 times larger than the median of the standard uncertainties associated with the 16 measured values of G.

Figure 8 reveals that the laboratory effects model entails generally smaller, more equitable increases to the effective uncertainties of the measured values than the common mean model, which involves multiplicative inflation of the reported uncertainties. Note that both panels of Figure 8 have the same scale in their vertical axes.

8.3 Evaluating Reproducibility

Table 5 summarizes the estimates of G and of other relevant quantities from Sections 8.1 and 8.2, alongside the CODATA 2018 recommended value of G and associated standard uncertainty [90]. These three estimates of G do not differ significantly from one another once their uncertainties are taken into account.

Schlamminger [82] notes that not only do "the various measurements of G seem not to converge on a value; it seems that the convergence gets worse with each additional data point." He concludes that "adding more data points from isolated experiments has not been the best strategy to improve the situation," and supports the idea of "forming an international consortium to coordinate these demanding experiments."

Such an international consortium [54] has meanwhile been formed, and in consequence the MARK-2 torsion balance that Quinn et al. [77, 78] built and used at the BIPM (International Bureau of Weights and Measures, Sèvres, France) was disassembled and shipped to NIST, in Gaithersburg, Maryland, U.S., where it was reassem-

TABLE 5

CODATA 2018 *recommended value of G* [90], *maximum likelihood estimate of G for the common mean model, and estimate of G from the Bayesian laboratory effects model described in Section* 8.2. *The corresponding standard uncertainties are listed under* $u(G)$. *The value for the dark uncertainty,* τ, *is the mean of its posterior distribution. The estimates of* κ, *which figures in equation* (2), *are the ad hoc estimate from Tiesinga et al.* [90], *and the maximum likelihood estimate. Only the latter is qualified with the associated standard uncertainty,* $u(\kappa)$

	G	$u(G)$	τ		
	$/(10^{-11}\ \mathrm{m^3kg^{-1}s^{-2}})$			κ	$u(\kappa)$
CODATA 2018	6.67430	0.00015		3.9	
Common Mean	6.67430	0.00013		3.5	0.6
Lab Effects	6.67399	0.00020	0.00096		

bled and mounted on a coordinate measurement machine; it became operational in August of 2016.

During the April 2022 meeting of the American Physical Society, Schlamminger et al. [83] described the new setup for the MARK-2 balance, explained how an independent, blind measurement of G was performed and announced that the first reproducibility test result should be revealed soon.

There are other avenues being explored to resolve this reproducibility crisis. One, merely data analytical, which in fact affords no resolution but only makes the consensus building more palatable, involves the use of a model with *shades of dark uncertainty*, which entertains not a single value of τ but several, which "penalize" different results differently [51]. Another, theoretical, employs non-classical physics models to explain the discrepancies between at least some of the historical results [40], and to "adjust" the affected results, thereby reducing the level of mutual inconsistency of the ensemble.

In brief, this review of the recent history of the measurement of the Newtonian constant of gravitation, G, and the corresponding quest for reproducibility provides yet another illustration of the extent to which the choice of statistical model (here between a common effect model and a laboratory effects model) impacts the assessment of reproducibility.

Maybe more importantly, it also shows that a reproducibility crisis can stimulate further research and encourage novel approaches to evaluate and improve reproducibility; in this case, the disassembly, transport across an ocean and reassembly at the destination of a delicate measuring instrument of great electromechanical complexity, as a radical and risky step taken on a wing and a prayer, hoping to identify reasons for the lack of reproducibility.

9. RECAPITULATION AND CONCLUSIONS

This contribution entertains a broad concept of reproducibility that is consistent with how this term has traditionally been understood in measurement science; the essential agreement between results when measuring the same property, or more generally studying the same phenomenon, while using different approaches, methods and procedures, applied by different experimenters working independently of one another in different laboratories and possibly at different times.

The illustrative examples show the key role that the evaluation of measurement uncertainty plays in identifying the seriousness of reproducibility crises, and in fleshing out, and quantifying, the impact that different causes, or sources of uncertainty, can have upon the lack of reproducibility.

The process of learning from experience through measurement is best done as a collective, collaborative enterprise, where different participants address the same problem and not only compare their results but also blend them into a consensus estimate. Such consensus estimate typically has smaller uncertainty than the uncertainty of the individual estimates taken separately, and is also supported by a richer, more varied basis of empirical evidence. The consensus estimate can be of interest in itself, as it is for the risk of rosiglitazone (Section 6) and for the reproduction number of a pandemic (Section 7), or it can provide a reference against which to compare individual measurement results, as it does in the measurement of G (Section 8).

The conclusions are most reliable when the methods variously employed by the participants are fundamentally different, possibly relying on different physical principles, and also when at least some of them are primary methods, in the sense explained in Section 3.2. In such cases, as Milton and Possolo [52] put it, "they achieve consilience" [94].

The precise nature of the aforementioned collective enterprise varies between meta-analysis in medicine and interlaboratory studies in measurement science. The former typically do not involve a preliminary agreement about methods and materials to be used by the participants, the onus of selecting the results to be compared and merged falling on the researcher conducting the meta-analysis. The latter usually are fairly structured procedures, involving a specified schedule and common protocols to be used for measurement.

The conventional understanding of reproducibility and repeatability in measurement science lends itself to the quantification of these attributes via some form or another of estimating variance components, as was illustrated for an interlaboratory study of the stress required to achieve a particular relative elongation of rubber samples (Section 5).

The examples also show that a meaningful data analysis can require a preliminary choice of reexpression for the measurement results, in particular to facilitate and legitimize the use of a statistical model that is demonstrably adequate for the data, and that is also fit for purpose. This was the case for the values of stress in the interlaboratory study of rubber elongation (Section 5), where a logarithmic reexpression was very helpful, and also for the meta-analysis for the effects of rosiglitazone (Section 6), with the traditional focus on log odds.

In interlaboratory studies and meta-analyses, there often arise results that deviate markedly from the bulk of the others; either because the measured value is rather different from most of the others, or because the uncertainty reported in a result is very different from the uncertainties reported in the other results, or both.

In general, and concerning very different reported uncertainties, it is the smallest uncertainties that are particularly influential, especially when the measurement results are mutually inconsistent, because they tend to pull the consensus value toward their corresponding measured values. Such unusually small uncertainties can then be said to be influential "inliers."

Faced with mutually inconsistent measurement results, the temptation is great to set "discrepant" values aside, thereby appearing to resolve the lack of reproducibility—Cox [24] describes one manner of succumbing to such temptation. However, unless there is a substantive, identifiable cause to do so, no "discrepant" result should be set aside, for the simple reason that in the absence of such cause there would be no logical basis whereon to reject discrepant values as being invalid—the most discrepant value can very well be the one closest to the true value of the measurand [26].

Statistical diagnostics are most valuable aids in identifying unusual measurement results, but statistical considerations alone are insufficient to reject a measurement result. Faced by challenges posed by "discrepant" but credible measurement results, one should tune the model to fit all credible results rather then set credible but "inconvenient" results aside. The example in Section 5 illustrated ways of accomplishing this, including by replacing the assumption that measurement errors are Gaussian with the assumption that they follow a Student's t-distribution with a small number of degrees of freedom, similar to [66].

The roller coaster that has been the history of the use of rosiglitazone as a therapy (Section 6) shows that, even when starting from the same set of data, one can reach rather different conclusions owing to different statistical models and methods of data reduction; in other words, the issue of lack of reproducibility raised its head when the results of alternative but comparably tenable models and data reductions were compared.

When blending independent estimates of the reproduction number for COVID-19 in the UK (Section 7), it so turned out that the mere exercise of preparing the inputs for analysis can be quite influential upon the level of reproducibility of the results, above and beyond the differences between the epidemiological models that provided those inputs, and also above and beyond the methods used to determine a consensus value. This serves as a warning about the fact that fairly simple matters often relegated to routine work can impact reproducibility, or the lack thereof, substantially.

The history of the measurements of the least accessible of the fundamental constants of nature, the Newtonian constant of gravitation, G, shows that alternative treatments of the same data, even when they produce results that are in fair agreement, involve very different assumptions that effectively establish dividing lines in the interested community; in particular, and in this case, whether one adopts the approach first proposed by Raymond Birge and faithfully followed mostly by the physics community, or opts instead for the approach that is prevalent in medical meta-analysis and in measurement science.

But the most important lesson one can draw from the recent history of the measurement of G is a lesson of optimism and empowerment; that, when faced with a considerable, genuine reproducibility crisis, the scientific community is ready to engage in extraordinary, cooperative efforts to understand the root causes of the lack of reproducibility, and to do so with the resolve needed to move heaven and earth, and with the creativity to match, of which Stephan Schlamminger (NIST) and his collaborators provide paradigmatic examples.

ACKNOWLEDGMENTS

The author is immensely grateful to Stefan Schlamminger (NIST) for all that he has taught him over the years about the measurement of G. The author is also much indebted to Olha Bodnar (Örebro University, Sweden), David Newton (NIST) and Mikela Waldman (NIST and Georgetown University, Washington, DC) for their most valuable and extensive suggestions for improvement of a draft of this contribution. The author thanks David Woods (Univ. of Southampton, UK) for an exchange of eMails about the measurement of the reproduction number of COVID-19 in the United Kingdom.

The author thanks the organizers of the special issue of *Statistical Science* dedicated to the issue of reproducibility for the invitation to contribute to it, and acknowledges the very helpful criticism and guidance that the guest editors, the journal's editor and a referee provided throughout the revision process, which led to considerable improvements.

Some specific commercial entities, equipment or materials may be identified in this document in order to describe or illustrate an experimental or statistical procedure or concept adequately. Such identification is not intended

to imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor is it intended to imply that the entities, equipment or materials mentioned are necessarily the best available for the purpose.

SUPPLEMENTARY MATERIAL

Data and R Code (DOI: 10.1214/23-STS899SUPP; .zip). The supplementary information file `Possolo 2023-TrackingTruth-Supplement.R` contains data and R code that facilitate reproducing the numerical results listed in this contribution.

REFERENCES

[1] AZZALINI, A. (2014). *The Skew-Normal and Related Families. Institute of Mathematical Statistics* (*IMS*) *Monographs* **3**. Cambridge Univ. Press, Cambridge. With the collaboration of Antonella Capitanio. MR3468021 https://doi.org/10.1017/cbo9781139248891

[2] BAKER, R. and JACKSON, D. (2015). New models for describing outliers in meta-analysis. *Res. Synth. Methods* **7** 314–328. https://doi.org/10.1002/jrsm.1191

[3] BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48. https://doi.org/10.18637/jss.v067.i01

[4] BEAUCHAMP, C. R., CAMARA, J. E., CARNEY, J., CHOQUETTE, S. J., COLE, K. D., DEROSE, P. C., DUEWER, D. L., EPSTEIN, M. S., KLINE, M. C. et al. (2021). *Metrological Tools for the Reference Materials and Reference Instruments of the NIST Materials Measurement Laboratory. NIST Special Publication* 260-136 (2021 *Edition*). National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.260-136-2021

[5] BELL, S. (1999). *A Beginner's Guide to Uncertainty of Measurement. Measurement Good Practice Guide* 11 (*Issue* 2). National Physical Laboratory, Teddington, Middlesex, United Kingdom. Amendments March 2001.

[6] BIPM (2019). *The International System of Units* (*SI*), 9th ed. International Bureau of Weights and Measures (BIPM), Sèvres, France.

[7] BIRGE, R. T. (1932). The calculation of errors by the method of least squares. *Phys. Rev.* **40** 207–227. https://doi.org/10.1103/PhysRev.40.207

[8] BLACKMAN, R. B. and TUKEY, J. W. (1958). The measurement of power spectra from the point of view of communications engineering. I. *Bell Syst. Tech. J.* **37** 185–282. MR0102897 https://doi.org/10.1002/j.1538-7305.1958.tb03874.x

[9] BLACKWELL, T., BROWN, C. and MOSTELLER, F. (1991). Which denominator? In *Fundamentals of Exploratory Analysis of Variance* (D. C. Hoaglin, F. Mosteller and J. W. Tukey, eds.) **10** 252–294. Wiley, New York, NY.

[10] BODNAR, O. and ELSTER, C. (2014). On the adjustment of inconsistent data using the Birge ratio. *Metrologia* **51** 516–521. https://doi.org/10.1088/0026-1394/51/5/516

[11] BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. T. and ROTHSTEIN, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* **1** 97–111. https://doi.org/10.1002/jrsm.12

[12] BRADBURN, M. J., DEEKS, J. J., BERLIN, J. A. and LOCALIO, A. R. (2007). Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Stat. Med.* **26** 53–77. MR2312699 https://doi.org/10.1002/sim.2528

[13] BÜRKNER, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80** 1–28. https://doi.org/10.18637/jss.v080.i01

[14] BÜRKNER, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* **10** 395–411. https://doi.org/10.32614/RJ-2018-017

[15] CAMPAGNARI, C. and MULDERS, M. (2022). An upset to the standard model. *Science* **376** 136–136. https://doi.org/10.1126/science.abm0101

[16] CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32. https://doi.org/10.18637/jss.v076.i01

[17] COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129. https://doi.org/10.2307/3001666

[18] ATLAS COLLABORATION, AABOUD, M. (2018). Measurement of the W-boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *European Physical Journal C* **78** 110. https://doi.org/10.1140/epjc/s10052-017-5475-4

[19] CDF COLLABORATION (2022). High-precision measurement of the W boson mass with the CDF II detector. *Science* **376** 170–176. https://doi.org/10.1126/science.abk1781

[20] L3 COLLABORATION (2006). Measurement of the mass and the width of the W boson at LEP. *Eur. Phys. J. C* **45** 569–587. https://doi.org/10.1140/epjc/s2005-02459-6

[21] ANALYTICAL METHODS COMMITTEE (1989a). Robust statistics—how not to reject outliers. Part 1. Basic concepts. *Analyst* **114** 1693–1697. https://doi.org/10.1039/AN9891401693

[22] ANALYTICAL METHODS COMMITTEE (1989b). Robust statistics—how not to reject outliers. Part 2. Inter-laboratory trials. *Analyst* **114** 1699–1702.

[23] COOPER, H., HEDGES, L. V. and VALENTINE, J. C., eds. (2019) *The Handbook of Research Synthesis and Meta-Analysis*, 3rd ed. Russell Sage Foundation Publications, New York, NY.

[24] COX, M. G. (2007). The evaluation of key comparison data: Determining the largest consistent subset. *Metrologia* **44** 187–200. https://doi.org/10.1088/0026-1394/44/3/005

[25] DAI, D. C. (2021). Variance of Newtonian constant from local gravitational acceleration measurements. *Phys. Rev. D* **103** 064059. https://doi.org/10.1103/PhysRevD.103.064059

[26] DE BIÈVRE, P. (2007). Statistics and measurement results in chemistry. *Accredit. Qual. Assur.* **12** 333–334. https://doi.org/10.1007/s00769-007-0294-1

[27] DELPHI COLLABORATION ABDALLAH, J. et al. Measurement of the mass and width of the W boson in e^+e^- collisions at $\sqrt{s} =$ 161-209 GeV. *Eur. Phys. J. C* **55** 1. https://doi.org/10.1140/epjc/s10052-008-0585-7

[28] DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

[29] DIAMOND, G. A., BAX, L. and KAUL, S. (2007). Uncertain effects of rosiglitazone on the risk for myocardial infarction and cardiovascular death. *Ann. Intern. Med.* **147** 578–581. https://doi.org/10.7326/0003-4819-147-8-200710160-00182

[30] FINEBERG, H. V., ALLISON, D. B., BARBA, L. A., CHONG, D., DONOHO, D., FREIRE, J., GABRIELSE, G., GATSONIS, C., HALL, E. et al. (2019). *Reproducibility and Replicability in Science. Committee on Reproducibility and Replicability in Science, the National Academies of Sciences, Engineering,*

and Medicine. The National Academies Press, Washington, DC. https://doi.org/10.17226/25303

[31] GAISER, C., FELLMUTH, B., HAFT, N., KUHN, A., THIELE-KRIVOI, B., ZANDT, T., FISCHER, J., JUSKO, O. and SABUGA, W. (2017). Final determination of the Boltzmann constant by dielectric-constant gas thermometry. *Metrologia* **54** 280–289. https://doi.org/10.1088/1681-7575/aa62e3

[32] PARTICLE DATA GROUP, ZYLA, P. A. et al. (2020). Review of Particle Physics. Progress of Theoretical and Experimental Physics 083C01. https://doi.org/10.1093/ptep/ptaa104

[33] GUNDERSEN, O. E. (2021). The fundamental principles of reproducibility. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **379** 20200210. https://doi.org/10.1098/rsta.2020.0210

[34] MICHELL, J. (2005). The logic of measurement: A realist overview. *Measurement* **38** 285–294. https://doi.org/10.1016/j.measurement.2005.09.004

[35] HARRIS, D. C. and LUCY, C. A. (2020). *Quantitative Chemical Analysis*, 10th ed. Macmillan Learning, New York, NY.

[36] HERSCHEL, J. F. W. (1866). Familiar Lectures on Scientific Subjects X. The Yard, the Pendulum, and the Metre 419–451, London Alexander Strahan.

[37] HOME, P. D., POCOCK, S. J., BECK-NIELSEN, H., CURTIS, P. S., GOMIS, R., HANEFELD, M., JONES, N. P., KOMAJDA, M. and MCMURRAY, J. J. V. (2009). Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): A multicentre, randomised, open-label trial. *Lancet* **373** 2125–2135. https://doi.org/10.1016/S0140-6736(09)60953-3

[38] JEWELL, N. P. (2004). *Statistics for Epidemiology*. CRC Press/CRC, Boca Raton, FL.

[39] KAHN, S. E., HAFFNER, S. M., HEISE, M. A., HERMAN, W. H., HOLMAN, R. R., JONES, N. P., KRAVITZ, B. G., LACHIN, J. M., O'NEILL, M. C. et al. (2006). Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. *N. Engl. J. Med.* **355** 2427–2443. https://doi.org/10.1056/NEJMoa066224

[40] KLEIN, N. (2020). Evidence for modified Newtonian dynamics from Cavendish-type gravitational constant experiments. *Classical Quantum Gravity* **37** 065002, 21. MR4086686 https://doi.org/10.1088/1361-6382/ab6cab

[41] KOEPKE, A., LAFARGE, T., POSSOLO, A. and TOMAN, B. (2017). Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia* **54** S34–S62. https://doi.org/10.1088/1681-7575/aa6c0e

[42] KOETSE, M. J., FLORAX, R. J. G. M. and DE GROOT, H. L. F. (2010). Consequences of effect size heterogeneity for meta-analysis: A Monte Carlo study. *Stat. Methods Appl.* **19** 217–236. MR2651450 https://doi.org/10.1007/s10260-009-0125-0

[43] LANGAN, D., HIGGINS, J. P. T., JACKSON, D., BOWDEN, J., VERONIKI, A. A., KONTOPANTELIS, E., VIECHTBAUER, W. and SIMMONDS, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res. Synth. Methods* **10** 83–98. https://doi.org/10.1002/jrsm.1316

[44] LANGAN, D., HIGGINS, J. P. T. and SIMMONDS, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Res. Synth. Methods* **8** 181–198. https://doi.org/10.1002/jrsm.1198

[45] MAISHMAN, T., SCHAAP, S., SILK, D. S., NEVITT, S. J., WOODS, D. C. and BOWMAN, V. E. (2022). Statistical methods used to combine the effective reproduction number, $R(t)$, and other related measures of COVID-19 in the UK. *Stat. Methods Med. Res.* **31** 1757–1777. MR4478307 https://doi.org/10.1177/09622802221109506

[46] MANDEL, J. (1972). Repeatability and reproducibility. *J. Qual. Technol.* **4** 74–85. https://doi.org/10.1080/00224065.1972.11980520

[47] MANDEL, J. (1991). The validation of measurement through interlaboratory studies. *Chemom. Intell. Lab. Syst.* **11** 109–119. https://doi.org/10.1016/0169-7439(91)80058-X

[48] MANDEL, J. and PAULE, R. (1970). Interlaboratory evaluation of a material with unequal numbers of replicates. *Anal. Chem.* **42** 1194–1197. https://doi.org/10.1021/ac60293a019

[49] MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22** 719–748. https://doi.org/10.1093/jnci/22.4.719

[50] MCCULLOCH, C. E., SEARLE, S. R. and NEUHAUS, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2431553

[51] MERKATAS, C., TOMAN, B., POSSOLO, A. and SCHLAMMINGER, S. (2019). Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation. *Metrologia* **56** 054001. https://doi.org/10.1088/1681-7575/ab3365

[52] MILTON, M. J. T. and POSSOLO, A. (2020). Trustworthy data underpin reproducible research. *Nat. Phys.* **16** 117–119. https://doi.org/10.1038/s41567-019-0780-5

[53] MISNER, C. W., THORNE, K. S. and WHEELER, J. A. (2017). *Gravitation*. Princeton University Press, Princeton, NJ.

[54] MOHR, P. (2014). Newtonian constant of gravitation international consortium. https://www.nist.gov/programs-projects/newtonian-constant-gravitation-international-consortium. NIST Physical Measurement Laboratory.

[55] MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2015). *CODATA Recommended Values of the Fundamental Physical Constants*: 2014. CODATA Zenodo Collection. https://doi.org/10.5281/zenodo.22826

[56] MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2016). CODATA recommended values of the fundamental physical constants: 2014. *Rev. Modern Phys.* **88** 035009. https://doi.org/10.1103/RevModPhys.88.035009

[57] MOLDOVER, M. R., TRUSLER, J. P. M., EDWARDS, T. J., MEHL, J. B. and DAVIS, R. S. (1988). Measurement of the universal gas constant R using a spherical acoustic resonator. *J. Res. Natl. Bur. Stand.* **93** 85–144. https://doi.org/10.6028/jres.093.010

[58] MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley Company, Reading, MA.

[59] MOULD, J. and UDDIN, S. A. (2014). Constraining a possible variation of G with type ia supernovae. *Publ. Astron. Soc. Austral.* **31** e015. https://doi.org/10.1017/pasa.2014.9

[60] MUNAFÒ, M. R., CHAMBERS, C., COLLINS, A., FORTUNATO, L. and MACLEOD, M. (2022). The Reproducibility Debate Is an Opportunity, Not a Crisis. BMC Research Notes 15 43. https://doi.org/10.1186/s13104-022-05942-3

[61] NEWELL, D. B. (2014). A more fundamental international system of units. *Phys. Today* **67** 35–41. https://doi.org/10.1063/PT.3.2448

[62] NISSEN, S. E. and WOLSKI, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N. Engl. J. Med.* **356** 2457–2471. https://doi.org/10.1056/NEJMoa072761

[63] NIST/SEMATECH (2012). *NIST/SEMATECH E-Handbook of Statistical Methods*. National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD. https://doi.org/10.18434/M32189

[64] NOZICK, R. (1981). *Philosophical Explanations*. Harvard Univ. Press, Cambridge, MA.

[65] OLVER, F. W. J., LOZIER, D. W., BOISVERT, R. F. and CLARK, C. W., eds. (2010) *NIST Handbook of Mathematical Functions*. Cambridge Univ. Press, Cambridge. MR2723248

[66] PINHEIRO, J. C., LIU, C. and WU, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. *J. Comput. Graph. Statist.* **10** 249–276. MR1939700 https://doi.org/10.1198/10618600152628059

[67] PINHEIRO, L. and EMSLIE, K. R. (2018). Basic concepts and validation of digital PCR measurements. In *Digital PCR*: *Methods and Protocols* 11–24 Springer, New York, New York, NY. https://doi.org/10.1007/978-1-4939-7778-9_2

[68] PLESSER, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Front. Neuroinform.* **11** 76. https://doi.org/10.3389/fninf.2017.00076

[69] PONTIUS, P. E. (1966). Measurement philosophy of the pilot program for mass calibration. National Bureau of Standards, Washington, DC. NBS Technical Note 288, Reprinted 1968, with minor corrections.

[70] POSSOLO, A. (2018). Measurement. In *Advanced Mathematical and Computational Tools in Metrology and Testing*: *AMCTM XI* (A. B. Forbes, N. F. Zhang, A. Chunovkina, S. Eichstädt and F. Pavese, eds.). *Series on Advances in Mathematics for Applied Sciences* **89** 273–285. World Scientific Company, Singapore. https://doi.org/10.1142/9789813274303\protect\T1\textunderscore0027

[71] POSSOLO, A. (2021). Concepts, methods, and tools enabling measurement quality. In *Frontiers in Statistical Quality Control* 13 (S. Knoth and W. Schmid, eds.) **19** 339–357. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-030-67856-2\protect\T1\textunderscore19

[72] POSSOLO, A. (2023). Supplement to "Tracking truth through measurement and the spyglass of statistics." https://doi.org/10.1214/23-STS899SUPP

[73] POSSOLO, A., BRUCE, S. S. and WATTERS, R. L. JR. (2021). *Metrological Traceability Frequently Asked Questions and NIST Policy*. National Institute of Standards and Technology, Gaithersburg, MD. NIST Technical Note 2156. https://doi.org/10.6028/NIST.TN.2156

[74] POSSOLO, A., KOEPKE, A., NEWTON, D. and WINCHESTER, M. R. (2021). Decision tree for key comparisons. *J. Res. Natl. Inst. Stand. Technol.* **126** 126007. https://doi.org/10.6028/jres.126.007

[75] POSSOLO, A. and MEIJA, J. (2022). *Measurement Uncertainty*: *A Reintroduction*, 2nd ed. Sistema Interamericano de Metrologia (SIM), Montevideo, Uruguay. https://doi.org/10.4224/1tqz-b038

[76] QU, J., BENZ, S. P., COAKLEY, K., ROGALLA, H., TEW, W. L., WHITE, R., ZHOU, K. and ZHOU, Z. (2017). An improved electronic determination of the Boltzmann constant by Johnson noise thermometry. *Metrologia* **54** 549–558. https://doi.org/10.1088/1681-7575/aa781e

[77] QUINN, T., PARKS, H., SPEAKE, C. and DAVIS, R. (2013). Improved determination of *G* using two methods. *Phys. Rev. Lett.* **111** 101102. https://doi.org/10.1103/PhysRevLett.111.101102

[78] QUINN, T., SPEAKE, C., PARKS, H. and DAVIS, R. (2014). The BIPM measurements of the Newtonian constant of gravitation, *G*. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **372** 0032. https://doi.org/10.1098/rsta.2014.0032

[79] ROUSH, S. (2005). *Tracking Truth*: *Knowledge*, *Evidence*, *and Science*. Oxford Univ. Press, New York, NY.

[80] RUKHIN, A. L. (2009). Weighted means statistics in interlaboratory studies. *Metrologia* **46** 323–331. https://doi.org/10.1088/0026-1394/46/3/021

[81] RUKHIN, A. L., BIGGERSTAFF, B. J. and VANGEL, M. G. (2000). Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *J. Statist. Plann. Inference* **83** 319–330. MR1748018 https://doi.org/10.1016/S0378-3758(99)00098-1

[82] SCHLAMMINGER, S. (2014). A cool way to measure big *G*. *Nature* **510** 478–480. https://doi.org/10.1038/nature13507

[83] SCHLAMMINGER, S., CHAO, L. S., LEE, V., SPEAKE, C. C. and NEWELL, D. B. (2022). Measurement of Newton's gravitational constant with the BIPM torsion balance. In *American Physical Society April Meeting* 2022 Session S16: Lab Experiments and Detector Characterization *S*16.00002.

[84] SCHLAMMINGER, S., HOLZSCHUH, E., KÜNDIG, W., NOLTING, F., PIXLEY, R. E., SCHURR, J. and STRAUMANN, U. (2006). Measurement of Newton's gravitational constant. *Phys. Rev. D* **74** 082001. https://doi.org/10.1103/PhysRevD.74.082001

[85] STRAIN, M. C., LADA, S. M., LUONG, T., ROUGHT, S. E., GIANELLA, S., TERRY, V. H., SPINA, C. A., WOELK, C. H. and RICHMAN, D. D. (2013). Highly precise measurement of HIV DNA by droplet digital PCR. *PLoS ONE* **8** 1–8. https://doi.org/10.1371/journal.pone.0055943

[86] R CORE TEAM (2022). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[87] STAN DEVELOPMENT TEAM (2022). RStan: the R interface to Stan. R package version 2.21.7.

[88] THOMAS, K. and SCHMIDT, M. S. (2012). Glaxo Agrees to Pay $3 Billion in Fraud Settlement. The New York Times July 2.

[89] THOMPSON, M. and ELLISON, S. L. R. (2011). Dark uncertainty. *Accredit. Qual. Assur.* **16** 483–487. https://doi.org/10.1007/s00769-011-0803-0

[90] TIESINGA, E., MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2021). CODATA recommended values of the fundamental physical constants: 2018. *Rev. Modern Phys.* **93** 025010. https://doi.org/10.1103/RevModPhys.93.025010

[91] VIBERTI, G., KAHN, S. E., GREENE, D. A., HERMAN, W. H., ZINMAN, B., HOLMAN, R. R., HAFFNER, S. M., LEVY, D., LACHIN, J. M. et al. (2002). A Diabetes Outcome Progression Trial (ADOPT): An international multicenter study of the comparative efficacy of rosiglitazone, glyburide, and metformin in recently diagnosed type 2 diabetes. *Diabetes Care* **25** 1737–1743. https://doi.org/10.2337/diacare.25.10.1737

[92] VIECHTBAUER, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36** 1–48. https://doi.org/10.18637/jss.v036.i03

[93] WHITE, R. (2011). The meaning of measurement in metrology. *Accredit. Qual. Assur.* **16** 31–41. https://doi.org/10.1007/s00769-010-0698-1

[94] WILSON, E. O. (1998). *Consilience*: *The Unity of Knowledge*. Alfred A. Knopf, New York, NY.

[95] YUSUF, S., PETO, R., LEWIS, J., COLLINS, R. and SLEIGHT, P. (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Prog. Cardiovasc. Dis.* **27** 335–371. https://doi.org/10.1016/s0033-0620(85)80003-7