Evaluation of the accuracy, consistency, and stability of measurements of the Planck constant used in the redefinition of the international system of units

Antonio Possolo, Stephan Schlamminger, Sara Stoudt, Jon R Pratt, and Carl J Williams

1 National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, United States of America
2 Department of Statistics, University of California, Berkeley, CA, United States of America

E-mail: antonio.possolo@nist.gov

Received 7 August 2017, revised 12 October 2017
Accepted for publication 27 October 2017
Published 12 December 2017

Abstract
The Consultative Committee for Mass and related quantities (CCM), of the International Committee for Weights and Measures (CIPM), has recently declared the readiness of the community to support the redefinition of the international system of units (SI) at the next meeting of the General Conference on Weights and Measures (CGPM) scheduled for November, 2018.

Such redefinition will replace the International Prototype of the Kilogram (IPK), as the definition and sole primary realization of the unit of mass, with a definition involving the Planck constant, $h$. This redefinition in terms of a fundamental constant of nature will enable widespread primary realizations not only of the kilogram but also of its multiples and sub-multiples, best to address the full range of practical needs in the measurement of mass.

We review and discuss the statistical models and statistical data reductions, uncertainty evaluations, and substantive arguments that support the verification of several technical preconditions for the redefinition that the CCM has established, and whose verification the CCM has affirmed. These conditions relate to the accuracy and mutual consistency of qualifying measurement results.

We review also an issue that has surfaced only recently, concerning the convergence toward a stable value, of the historical values that the Task Group on Fundamental Constants of the committee on Data for Science and Technology (CODATA-TGFC) has recommended for $h$ over the years, even though the CCM has not deemed this issue to be relevant.

We conclude that no statistically significant trend can be substantiated for these recommended values, but note that cumulative consensus values that may be derived from the historical measurement results for $h$ seem to have converged while continuing to exhibit fluctuations that are typical of a process in statistical control.

Finally, we argue that the most recent consensus value derived from the best measurements available for $h$, obtained using either a Kibble balance or the $xrcd$ method, is reliable and has uncertainty no larger than the uncertainties surrounding the current primary and secondary realizations of the unit of mass, hence that no credible technical impediments stand in the way of the redefinition of the unit of mass in terms of a fixed value of $h$. 

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
1. Introduction

In May of 2017, the Consultative Committee for Mass and related quantities (CCM) of the International Committee for Weights and Measures (CIPM) met at the International Bureau of Weights and Measures (BIPM) in Sèvres, France, and recommended ‘that the CIPM undertakes the necessary steps to proceed with the planned redefinition of the SI at the next meeting of the CGPM, acknowledging the measures to be taken by the CCM to ensure integrity and continuity in the dissemination of the kilogram’.

The CCM reached this conclusion based on an evaluation of whether several prerequisites, agreed upon previously, had been met by the metrological community represented in the CCM. These requirements included considerations of individual measurement accuracy, and homogeneity (or mutual consistency) of methodologically diverse, sufficiently accurate measurement results for the Planck constant, \( h \). ‘Homogeneity’ means that the dispersion of the measured values is not excessive in light of the reported measurement uncertainties.

In addition, the CCM also considered an issue that has surfaced only recently, and that is not part of the preconditions for recommending a redefinition of the unit of mass: the convergence toward a stable value of the consensus values that the CODATA-TGFC has recommended over the years.

The question has also been discussed in several venues, of whether the uncertainty surrounding the estimate of \( h \), at the time when its value will be fixed, will translate into an uncertainty in the realization of the unit of mass that is fit for the purposes of the vast and diverse applications where mass is measured.

This contribution reviews and discusses statistical models and methods for data reduction that support the aforementioned recommendation issued by the CCM, and that otherwise contribute to elucidate the related issues just mentioned.

In section 2, we present a historical retrospective of the measurements made of the Planck constant since 1979, and also consider the issue of whether there is any statistically significant temporal trend in the values of \( h \) that the CODATA-TGFC has recommended over the past 30 years.

In section 3, we evaluate the accuracy requirements defined by the CCM and identify the subset of published measurement results that satisfy these requirements. Section 4 examines the issue of mutual consistency of select subsets of the set of measurement results that meet the accuracy requirements, from both statistical and substantive viewpoints.

Finally, section 5 recapitulates the more salient findings, highlights the accomplishments that justify the redefinition of the kilogram in terms of the Planck constant, sketches anticipated research and development following the redefinition, and expresses unequivocal support for the recommendation from the CCM in favor of the much and long anticipated redefinition.

2. Retrospective

Table 1 lists measurement results for \( h \) that either are listed in table 21 of [38], or that otherwise qualify for consideration by the CODATA-TGFC in relation with the special adjustment that will produce the values of the fundamental constants recommended for use in the redefinition of the SI.

Figure 1 represents the historical, cumulative evolution of the consensus value for \( h \) derived from the results in table 1. For example, the (blue) dot corresponding to 1979 represents only the value measured by NPL-79. But the dot above 1980 represents the consensus value that blends the results from both NPL-79 and NIST-80. Similarly for all the other points. The last point on the right side of the plot, above the last instance of 2017, is the consensus value that was derived from all the data in that table.

The 2014 CODATA recommended value of \( h \), and the associated uncertainty, both depicted in the same figure, serve only as a reference and basis for comparison. It is an interesting coincidence that the sequence of cumulative consensus values gravitates toward it, and indeed have been statistically indistinguishable from it since 2011 at least. The variability of the cumulative consensus values in the more recent past, say since 2015, is very much what should be expected for a process in statistical control [9].

The retrospective suggests readiness for fixing the value of \( h \) somewhere within the range of the uncertainty of the final, most current and comprehensive consensus value that the CODATA-TGFC will derive from measurement results that include or are closely related to those listed in table 1. The preliminary value, \( 6.626\,070\,150(69) \times 10^{-34} \, \text{J s} \) [39], that the CODATA-TGFC will recommend for \( h \) as a result of the special adjustment of 2017 is depicted in figure 2.

The consensus values depicted in figure 1 could have been computed using any one of many, generally accepted methods available for consensus building. We have used the most common procedure in the area of application (meta-analysis in medicine) where most inter-comparisons are made, the so-called DerSimonian–Laird procedure [18]. The procedure is implemented in package metafor for the R environment for statistical computing and graphics [41, 47], and is readily available in a web-based application with a user-friendly interface, the NIST Consensus Builder, which is available worldwide via any common Web browser, at https://consensus.nist.gov [28, 29].
The statistical model underlying the DerSimonian–Laird procedure is an additive, random effects model, which represents each measured value as a sum of three elements,

\[ h_j = h + \lambda_j + \epsilon_j, \]

for \( j = 1, \ldots, n \) different experiments, where \( h \) denotes the measurand, the \( \{\lambda_j\} \) denote experiment (or laboratory) effects, and the \( \{\epsilon_j\} \) represent experiment-specific measurement errors.

The \( \{\epsilon_j\} \) are regarded as realized values of independent random variables with mean 0 and possibly different standard deviations assumed equal to the standard uncertainties reported in the original publications. The \( \{\lambda_j\} \) are modeled as values of independent random variables with mean 0 and standard deviation \( \tau \geq 0 \). These assumptions imply that, taken collectively, the measured values are centered at the true value of \( h \).

The standard deviation \( \tau \) of the experiment (or laboratory) effects \( \{\lambda_j\} \) quantifies the ‘excess’ dispersion of the measured values, above and beyond what should be expected given the reported uncertainties. Because it remains ‘invisible’ until the measurement results are inter-compared, it is often called dark uncertainty [46].

An alternative model that is often used by the CODATA–TGFC, originally suggested by [7] for the combination of heterogeneous (or mutually inconsistent) measurement results, represents the measured values as \( h_j = h + \kappa \epsilon_j \), where now the true value of \( h \) is a common mean obtained after inflating the reported uncertainties by a multiplicative adjustment \( \kappa \), typically greater than 1. [2, 29] compare and discuss the additive and multiplicative models at considerable

### Table 1

The 13 measurement results labeled NPL-79 to NIST-15 are inferred values of the Planck constant compiled in table 21 of [38], who list all the underlying references. The result for NRC-15, which that same table also lists, has meanwhile been superseded by NRC-17. The labels in the third column indicate whether the measurement was made using a Kibble (or watt) balance (k), a Joule balance (j), some other electrical method (e), or the x-ray silicon crystal density (XRCd) method (x).

| Experiment | \( h \times 10^{-34} \text{ J s} \) | \( u(h) \times 10^{-41} \text{ J s} \) | \( u(h)/h \times 10^{-8} \) |
|------------|----------------|----------------|----------------|
| NPL-79     | 6.626 073      | 67             | E              |
| IAC-11     | 6.626 06989    | 2              | 3              | X              |
| METAS-11   | 6.626 0691     | 20             | 30             | K              |
| NPL-12     | 6.626 0712     | 13             | 20             | K              |
| IAC-15     | 6.626 07015    | 1.3            | 2              | X              |
| LNE-15     | 6.626 0688     | 17             | 2              | K              |
| NIST-15    | 6.626 06936    | 3.8            | 6              | K              |
| NRC-17     | 6.626 07013    | 0.6            | 1              | 1              | Wood et al [48] |
| LNE-17     | 6.626 07041    | 3.8            | 6              | 1              | Thomas et al [45] |
| NMIJ-17    | 6.626 070059   | 1.65           | 2              | X              | Kuramoto et al [30] |
| NIM-17     | 6.626 0692     | 16             | 24             | J              | Li et al [31] |
| IAC-17     | 6.626 070404   | 0.792          | 1              | X              | Bartl et al [3] |
| NIST-17    | 6.626 069934   | 0.89           | 1              | K              | Haddad et al [23] |

* Indicate the six experiments whose relative standard uncertainty is no larger than \( 5 \times 10^8 \).

The historical evolution of the consensus value for \( h \). Each (blue) dot represents the consensus value derived from all previous measurement results. The horizontal (red) line indicates the 2014 CODATA recommended value of \( h \), and serves only as a reference and basis for comparison.
length. We will discuss the Birge adjustment further in section 4.2.

2.1. Thirty years of CODATA recommendations

The question has been raised of how the values that the CODATA-TGFC has been recommending for $h$ over the years have evolved over time, and whether any trend such evolution might display could possibly challenge the readiness of the state-of-the-art to support the redefinition of the kilogram.

Neither was the absence of a trend one of the requirements originally specified by the CCM [22], nor was it a primary consideration issuing from the CCM meeting of May 2017. But since it was discussed during this meeting as a potential constraint on the date for the redefinition, we review it here taking into account the sources of uncertainty that affect any apparent trend.

It should be noted that values that the CODATA-TGFC has recommended for $h$ in its periodic adjustments are not based on a consensus building exercise as we have undertaken above, but result instead from a least-squares adjustment that simultaneously involves all of the qualifying measurement results that are informative about all the fundamental constants (except the Newtonian constant of gravitation, $G$, which is treated separately), possibly after applying multiplicative adjustments to reported uncertainties that the CODATA-TGFC deems to be problematic (typically for appearing to be too small).

This is the reason why the caption of table 21 of [38] explains that the values listed for $h$ are *inferred values* obtained from the experimental data considered in the overall data reduction, and similarly for corresponding tabulations for other years.

Figure 2 depicts the values that the CODATA-TGFC has recommended for $h$ in the course of the past 30 years (selected from the values listed in table 2), the preliminary CODATA 2017 value for $h$ [39] (open blue circle), a suggested ‘trend’ (red curve), and 99% (pink region) coverage bands for what the true trend could possibly be.

The fact that the uncertainty bands can fully accommodate a horizontal line segment drawn across the whole range of dates represented on the horizontal axis, shows that the evidence is insufficient to reject the hypothesis of there being no trend. The relative value at 2017, of the slope (first derivative) of the red curve, equals 2 parts in $10^9$ per year, with standard uncertainty 6 parts in $10^9$ per year, hence does not differ significantly from 0.

The coverage bands take into account (a) the uncertainties associated with the CODATA recommended values, (b) the correlations between the values recommended in different releases, and (c) the uncertainties associated with these correlations.

The contribution from the correlations mentioned in (b) is very important, and any evaluation of the significance of any
presumptive trend that ignores it, will be unrealistic. These correlations range from 0 to 0.73, and arise because different releases tend to share substantial proportions of the same experimental results.

We have used (1) one instance of the Monte Carlo method to estimate the correlations mentioned in (b), and (2) another instance to propagate the contributions from (a)–(c) and to evaluate the uncertainty associated with the aforementioned trend.

(1) We explain the first instance of the Monte Carlo method for the correlation between the values CODATA recommended in 2010 and 2014, the others having been estimated similarly. The value for 2014 subsumes 14 determinations of h, of which 10 were also considered in the 2010 release. The same measurement errors affecting these 10 shared values are present in both the 2010 and 2014 releases, hence the correlation (which turned out to be 0.54) between the corresponding consensus values, which we computed using the aforementioned DerSimonian–Laird procedure.

To estimate this correlation we applied perturbations to the measured values repeatedly, commensurate with the associated uncertainties, with the same perturbations being applied to the same measured values that the CODATA–TGFC used in adjustments for 2010 and 2014, and then estimated the correlation using all simulated pairs of adjusted values for these two years.

(2) The second instance of the Monte Carlo method was used to propagate the contributions that (a)–(c) make to the uncertainty associated with the ‘trend’ depicted in figure 2. This ‘trend’ was defined by fitting a spline to the CODATA recommended values using a generalized additive model as implemented in R function gam, defined in package mgcv [49]. The Monte Carlo method used was a version of the so-called parametric statistical bootstrap [19], and involved two stages of sampling, as follows:

(2.1) First, by sampling each correlation from the approximately Gaussian distribution of its Fisher’s z-transform [20, 21]. This recognizes and takes into account the fact that each of the correlations, although estimated based on many pairs of consensus values of h, in fact are based on a fairly small number of values shared by the adjusted values of h that correspond to two different years.

(2.2) Second, by drawing from a multivariate Gaussian distribution whose covariance matrix reflects both the uncertainties associated with the values of h recommended in different years, and the correlations between them.

Since the correlation matrix produced in (2.1) need not be positive-definite, it was replaced in each iteration of the parametric bootstrap procedure by the ‘closest’ positive-definite correlation matrix computed using R function nearPD defined in the package Matrix [4, 25].

The coverage band is based on 50000 samples drawn from the multivariate distribution of the consensus values of h corresponding to the six different years under consideration, and was computed using methods described by [16] and implemented in R function boot envolope defined in package boot [10].

The band has been constructed to reflect the uncertainty associated with the ‘true’ overall trend over the whole range of times represented along the horizontal axis, therefore at all times simultaneously, rather than at each instant in time without controlling coverage at other times. Such a band performs is wider than the band intended to achieve only point-wise coverage, but it is the appropriate one to gauge the reality of the ‘trend’ as a whole. Even cursory examination of either version of this simultaneous coverage band (one for 95% confidence, the other for 99% confidence) reveals that no trend can properly be substantiated because either band can accommodate a horizontal line, signifying no trend.

The substantively relevant questions relating to the values the CODATA–TGFC has recommended over the years are about (i) the relative uncertainty that the CODATA–TGFC will use to qualify the value recommended for h in the special adjustment of 2017, and about (ii) whether this relative uncertainty warrants the redefinition in the sense that it guarantees seamless continuity in the primary realizations of the unit of mass before and after the redefinition, to within an uncertainty that is comparable to the uncertainties prevailing in the current, routine dissemination of the IPK [1].

The preliminary value of the answer to (i) is 10 parts in 10^8 [39], and the discussion in the following sections will only reinforce the view that (ii) should be answered in the affirmative.

3. Accuracy

CCM recommendation g1, of 2017, for a new definition of the kilogram in 2018, notes that the accuracy requirement has been met. The requirement specifies that ‘at least three independent experiments, including work from Kibble balance and xrcd experiments, yield values of the Planck constant with relative standard uncertainties not larger than 5 \times 10^{-8}, and that ‘at least one of these results should have a relative standard uncertainty not larger than 2 \times 10^{-8}’ [14].

In fact there are now six experiments with relative standard uncertainties smaller than 5 \times 10^{-8}, marked with a superscript (a) in table 1: IAC-11, IAC-15, NRC-17, NMIIJ-17, IAC-17, and NIST-17. Four of these have relative standard uncertainties smaller than 2 \times 10^{-8}: IAC-15, NRC-17, IAC-17, and NIST-17.

Furthermore, the Pilot Study CCM.R-kg-P1 [43], which involved the BIPM, LNE (France), NIST (USA), NMII (Japan), NRC (Canada), and PTB (Germany), provided a compelling demonstration of the mutual consistency that can be achieved by multiple primary realizations of the unit of mass. The weighted mean of the measurement results obtained by the participants agreed with the calibration based on the IPK to
within 0.001 mg for one set of 1 kg standards used in the comparison, and to within 0.0045 mg for the other.

4. Consistency

4.1. Cochran’s test

The CCCM recommendation includes the observation that the ‘most recent measurement results with relative standard uncertainty below $5 \times 10^3$ do not pass the standard chi-squared test of consistency’. This test, which is commonly referred to as Cochran’s $Q$ test [11], is often used to detect whether measurement results are heterogeneous (or mutually inconsistent) in the sense that the measured values are more dispersed than their reported uncertainties would have led one to expect.

The statistic employed in this test is $Q = \sum_{j=1}^{n} (h_j - \hat{h})^2 / w_j^2$, where $\hat{h}$ denotes the consensus value, $h_1, \ldots, h_n$ are the estimates of $h$ produced in the $n$ experiments, and $w_1, \ldots, w_n$ are the reported measurement uncertainties (which are interpreted as the standard deviations of the measurement errors $\{\varepsilon_j\}$ in the models discussed in section 2). In all cases where we use Cochran’s test, $\hat{h}$ denotes the weighted average $\sum_{j=1}^{n} w_j h_j / \sum_{j=1}^{n} w_j$, with $w_j = 1 / w_j^2$ for $j = 1, \ldots, n$.

For the $n = 6$ measurement results marked with the superscript (*) in table 1, $Q = 18.3$. The reference probability distribution on the assumption of homogeneity is chi-squared with $n - 1 = 5$ degrees of freedom. Hence the test has $p$-value 0.003, which is the probability of observing a value of $Q$ at least as large as the one that was observed, owing to the vagaries of sampling alone, when the results truly are homogeneous, or mutually consistent. Since this probability is quite small, the conventional conclusion is that the results are heterogeneous (or mutually inconsistent).

The test suffers from well-known, and potentially consequential limitations [26]. The more relevant limitations from the viewpoint of the present discussion are: first, when many measurement results are involved, the test may detect statistically significant heterogeneity that is substantively irrelevant [24]; second, the validity of the test hinges on the assumption that the $\{w_j\}$ are all based on infinitely many degrees of freedom, which cannot possibly be the case here. Therefore, the conclusion of the test must be regarded with circumspection, and ought not to be taken as definitive.

Alternative tests of mutual consistency, recently suggested by [33], lead to different conclusions. For example, the test based on $Q = \sum_{j=1}^{n} |h_j - \hat{h}| / w_j$, where $\hat{h}$ is the same weighted average specified above, when applied to the 6 measurement results aforementioned, has $p$-value 0.022. And if $\hat{h}$ is chosen as the weighted median instead, then the $p$-value increases to 0.18. Both $p$-values just quoted were obtained by application of a bootstrap procedure described by [33, section 3.3] and implemented in R function metahet defined in package altmeta [32].

For the purposes of the redefinition, however, much more important than the appearance of heterogeneity as judged by a statistical test of questionable validity, is the actual impact that the apparent heterogeneity, if indeed it were to be substantiated, might have upon the measurement of mass based on a consensus value of $h$ derived from these measurement results.

To evaluate this impact, let us reason as follows: if the measurement results indeed are heterogeneous (or mutually inconsistent), and one computes the consensus value six times over, each time leaving out one of the results in turn, then the dispersion of the six leave-one-out estimates will express the substantive impact of the alleged heterogeneity.

The leave-one-out weighted averages (divided by $10^{-34}$ J s) are 6.62607016 (IAC-11), 6.62607015 (IAC-15), 6.62607016 (NRC-17), 6.62607015 (NMIJ-17), 6.62607007 (IAC-17), and 6.62607020 (NIST-17), where the labels between parenthesis indicate the experiments whose measurement results have been left out of the weighted averages. Their sample standard deviation (again divided by $1 \times 10^{-4}$ J s) is $4.2 \times 10^{-8}$, which amounts to 6 parts in $10^9$ of the weighted average of all six measured values in this set.

Considering that, according to the OIML R111-1 standard [27], the maximum permissible error for a class E2 standard of mass 1 kg is 0.5 mg, which translates into a relative standard uncertainty of $((0.5/3)/2) \, \text{mg} \, \text{kg}^{-1} = 83$ parts in $10^9$ [1], those 6 parts in $10^9$ pale by comparison, thus suggesting that the statistically significant heterogeneity is substantively negligible.

Even the relative standard deviation of the six, qualifying measured values is less than 30 parts in $10^9$, which implies that they are exchangeable in practice, in the sense that any one of them could be selected to serve as the sole primary realization of the kilogram instead of a consensus value, up to this level of relative uncertainty.

This relative dispersion of 30 parts in $10^9$ is also smaller than the 35 $\mu$g shift that was applied to the BIPM ‘as-maintained mass unit’, following the Extraordinary Calibrations of 2013–2014 [42]. This calibration campaign, which involved comparing the ten working standards that the BIPM uses routinely for the dissemination of the unit of mass, against the IPK, produced the surprising finding that, during the period 1992–2014, all BIPM working standards had lost mass with respect to the IPK, by 35 $\mu$g on average [17].

4.2. Birge’s adjustment

It is common practice for the CODATA-TGFC to apply a sufficiently large multiplicative correction, to the measurement uncertainties in results that are heterogeneous, to achieve consistency, based on the belief that the reported uncertainties generally underestimate the values that they should have [7]. The same goal can be achieved in the context of the additive random effects model mentioned in section 2, by introducing experiment or laboratory effects $\{\lambda_j\}$ with expected value 0 and standard deviation $\tau$ (dark uncertainty).

The conventional implementation of the multiplicative adjustment sets the adjustment factor $\kappa$ equal to the so-called Birge ratio, which is $\sqrt{Q/(n-1)} = 1.9$ for the six, selected measurement results. The same metric has been ‘rediscovered’ in the field of meta-analysis in medicine, where it is called the $H$ statistic [24].
First, it should be noted that the Birge ratio, being a function of the measured values, has an associated uncertainty, which qualifies the extent of the heterogeneity (or mutual inconsistency) of the measurement results. Higgins and Thompson describe eight different methods that may be used to evaluate this uncertainty, and [8] describe yet another. For the six measurement results marked with the superscript (a) in table 1, a Monte Carlo method implemented in the aforementioned R function metahet produces a 95% confidence interval for the Birge ratio ranging from 0.51 to 2.55. Since the interval includes 1, which is its expected value in the absence of heterogeneity, no correction appears to be warranted.

Second, a value of $\kappa$ appreciably smaller than that 1.9 suffices to produce a value of the test statistic $Q$ that is not statistically significant in a test with probability 0.05 of erroneously rejecting the hypothesis of homogeneity when the hypothesis is true. In this case, the value is $\hat{\kappa} = 1.285$.

Now, an increase by a mere 28.5% of the uncertainties reported in those six experiments, which makes their results mutually consistent as judged by Cochran’s test, is much smaller than similar increases that the CODATA-TGFC has applied to produce recommended values for several of the fundamental constants in past data reductions. Furthermore, even after multiplying the reported uncertainties by 1.285, the accuracy criteria set by the CCM continue to be satisfied by the same set of six measurement results marked with the superscript (a) in table 1.

Interestingly, the maximum likelihood estimate of $\kappa$, for the multiplicative model $h_j = h + \kappa \epsilon_j$, under the additional assumptions that the $\{\epsilon_j\}$ are independent and Gaussian, is $\hat{\kappa} = 1.745$, which lies between what the Birge ratio suggests and the minimal value for which Cochran’s test no longer rejects the hypothesis of homogeneity.

Furthermore, application of the parametric statistical bootstrap to evaluate the uncertainty associated with $\hat{\kappa}$ produces a 95% confidence interval for it ranging from 0.65 to 2.50 [19]. Therefore, and similarly to what was noted above, since this interval includes 1, from the viewpoint of the multiplicative model there is no compelling statistical reason for expanding the reported uncertainties, and the six results effectively appear to be mutually consistent.

### 4.3. Homogeneous triplets

Considering the two accuracy requirements together that the CCM agreed should be met to warrant the redefinition of the kilogram in 2018—at least three measurements with relative standard uncertainties no larger than $5 \times 10^8$, and at least one of these not larger than $2 \times 10^8$—leads naturally to the question of whether any such triplets are homogeneous (or mutually consistent).

The set of six experiments with relative standard uncertainties not larger than $5 \times 10^8$ has 20 subsets of size 3. Of these, nine subsets are homogeneous and comprise at least one experiment whose relative standard uncertainty is no larger than $2 \times 10^8$, at least one Kibble balance experiment, and at least one experiment using the x-ray silicon crystal density (xrcd) method. Homogeneity is taken to mean that the hypothesis of mutual consistency is not rejected in Cochran’s $Q$ test with probability 0.05 of erroneous rejection when the hypothesis is true. Table 3 lists the compositions of these nine subsets of size 3. Similarly, there are five homogeneous subsets of size 4 (out of 15) that meet the same requirements.

### 5. Conclusions

The level of agreement between the most accurate determinations of $h$ available today is simply remarkable, especially considering that they involve multiple, different, and independent versions of two radically different approaches to its measurement.

Furthermore, both the Kibble balance and the xrcd method successfully provide an absolute realization of the unit of mass, a fact that has not been attempted since the national convention of the First Republic of France passed a decree on April 7, 1795, defining the gram as the ‘absolute weight of a volume of pure water equal to the cube of the one hundredth part of a meter, at the temperature of melting ice’.

This initial definition of the unit of mass would have made primary realizations of it widely accessible in principle, notwithstanding the technical difficulties of implementing it in practice. However, as soon as a choice was made literally to cast the unit of mass into a fixed, physical artifact, first as the Kilogram of the Archives, later as the international prototype of the kilogram (IPK), the measurement of mass has involved only comparisons with standards linked to a single primary realization through a traceability chain, and no absolute determinations [40].

The investments that the national metrology institutes have committed to the realization of the unit of mass have finally paid off. The accuracy and methodological diversity requirements that the CCM had established as preconditions for the redefinition of the unit of mass, and for its realization in terms of a fundamental constant of nature, have been met.

The residual heterogeneity of the most accurate measurements of $h$ is too small to be a cause for concern in practice. In fact, the relative uncertainty in the consensus value of $h$ that may be derived from the six measurement results that currently satisfy the stringent accuracy requirements established by the CCM, and that is attributable to residual heterogeneity, amounts to 6 parts in $10^9$. 

### Table 3. Each row defines the composition of a triplet of experiments that produced homogeneous results, all of which have relative standard uncertainties no larger than $5 \times 10^8$, at least one of them being no larger than $2 \times 10^8$, and at least one used a Kibble balance and another the xrcd method. The last column lists the $p$-values of Cochran’s $Q$ test of homogeneity.

| Composition | $p$-value |
|-------------|-----------|
| {IAC-11, IAC-15, NRC-17} | 0.49 |
| {IAC-11, IAC-15, NIST-17} | 0.34 |
| {IAC-11, NRC-17, NMIJ-17} | 0.48 |
| {IAC-11, NRC-17, NIST-17} | 0.12 |
| {IAC-11, NMIJ-17, NIST-17} | 0.76 |
| {IAC-15, NRC-17, NMIJ-17} | 0.90 |
| {IAC-15, NMIJ-17, NIST-17} | 0.15 |
| {IAC-15, NMIJ-17, NIST-17} | 0.37 |
| {NRC-17, NMIJ-17, NIST-17} | 0.18 |
This is practically negligible considering that the relative standard uncertainty allowed for the most reliable 1 kg mass standards recognized by the international organization of legal metrology (OIML) is 83 parts in 10^9.

As an additional term of comparison, we point out that, in 2012, the BIPM assigned standard uncertainty 0.007 mg (that is, 7 parts in 10^9) to the 1 kg mass prototype number 85, belonging to the United States of America (BIPM Certificate No. 30, 2012, signed by M Kühne, Director). Other standards, calibrated at approximately the same time, have received either the same or closely similar uncertainty qualifications associated with their assigned masses.

Even if a single one of those six best determinations were to be randomly selected and assigned to h as its final and fixed value, the corresponding ‘cost’ that would be incurred in relative definitional uncertainty would be no more than 30 parts in 10^9, which is still less than the steadily increasing, historical dispersion of the masses of the IPK, of its official copies, and of the national prototypes, relative to one another [15, 40]. That uncertainty would also be smaller than the shift of 35 μg that was applied to the BIPM ‘as-maintained mass unit’, following the Extraordinary Calibrations of 2013–2014.

The collective effort that enabled the current accomplishments is far ranging because it brings to within the reach of many laboratories, not just national metrology institutes—national and local government agencies, universities, and industry—the ability to realize the unit of mass. It will also facilitate and reduce the time and costs to disseminate this unit throughout what now effectively becomes a network of traceability chains anchored on multiple primary realizations, increasing the reliability of mass measurements all the way down to the ultimate consumer of such measurements.

Being a collective effort means that no participant has been left out or behind, and that all have added tangible value to the enterprise, regardless of how any particular measurement result may be situated relative to the consensus value that blends them all and indeed enables a reliable redefinition of the kilogram.

Neither does reaching this milestone imply that the research and development that has brought us to this point need not continue. Quite the contrary. Once the value of the Planck constant is fixed, the uncertainty surrounding the consensus value that will be assigned to h does not evaporate: instead, it should be converted into a common definitional, or realization uncertainty, that percolates to all instances of the realization. Continued research and development, and improvement of the measurement methods, most likely will reduce this uncertainty component going forward, hence will steadily improve performance in the measurement of mass.

Acknowledgments

The authors are much and variously indebted to Zeina Kubarych (NIST) and Jolene Splett (NIST) for the numerous and most valuable suggestions for improvement that they offered. David Newell (NIST, and chair of the CODATA-TGFC) kindly shared the preliminary value recommended for h resulting from the special adjustment of 2017, and pointed out the absolute nature of the unit of mass adopted by the National Convention of the First Republic of France. Sara Stoldt is supported by a National Physical Science Consortium fellowship.

ORCID iDs

Antonio Possolo https://orcid.org/0000-0002-8691-4190
Stephan Schlamminger https://orcid.org/0000-0002-9270-4018

References

[1] Abbott P J and Kubarych Z J 2013 The new kilogram definition and its implications for high-precision mass tolerance classes J. Res. Natl Inst. Stand. Technol. 118 353–8
[2] Baker R D and Jackson D 2013 Meta-analysis inside and outside particle physics: two traditions that should converge? Res. Synth. Methods 4 109–24
[3] Bartl G et al 2017 A new 28Si single crystal: counting the atoms for the new kilogram definition Metrologia 54 693
[4] Bates D and Maechler M 2017 Matrix: sparse and dense matrix classes and methods https://CRAN.R-project.org/package=Matrix R package version 1.2–10
[5] Birge R T 1919 The most probable value of the Planck constant h Phys. Rev. 14 361–8
[6] Birge R T 1929 Probable values of the general physical constants Rev. Mod. Phys. 1 1–73
[7] Birge R T 1932 The calculation of errors by the method of least squares Phys. Rev. 40 207–27
[8] Bodnar O and Elster C 2014 On the adjustment of inconsistent data using the Birge ratio Metrologia 51 516–21
[9] Box G E P and Luceño A 1997 Statistical Control by Monitoring and Feedback Adjustment (New York: Wiley)
[10] Canty A and Ripley B 2017 Boot: bootstrap R (S-Plus) functions cran.r-project.org/web/packages/boot/ R package version 1.3-19
[11] Cochran W G 1954 The combination of estimates from different experiments Biometrics 10 101–29
[12] Cohen E R and Taylor B N 1973 The 1973 least-squares adjustment of the fundamental constants J. Phys. Chem. Ref. Data 2 663–734
[13] Cohen E R and Taylor B N 1987 The 1986 adjustment of the fundamental physical constants Rev. Mod. Phys. 59 1121–48
[14] Consultative Committee for Mass (CCM). Recommendation of the Consultative Committee for Mass and related quantities submitted to the International Committee for Weights and Measures, BIPM, Sèvres, France, May 2017. www.bipm.org/cc/CCM/Allowed/1606E_Final_CCM-Recommendation_2017.pdf, Recommendation G1, for a new definition of the kilogram in 2018
[15] Davis R 2003 The SI unit of mass Metrologia 40 299–305
[16] Davison A C and Hinkley D 1997 Bootstrap Methods and their Applications (New York: Cambridge University Press)
[17] deMirandès E, Barat P, Stock M and Milton M J T 2016 Calibration campaign against the international prototype of the kilogram in anticipation of the redefinition of the kilogram, part II: evolution of the BIPM as-maintained mass unit from the 3rd periodic verification to 2014 Metrologia 53 1204–14
[18] DerSimonian R and Laird N 1986 Meta-analysis in clinical trials Control. Clin. Trials 7 177–88
[19] Efron B and Tibshirani R J 1993 An Introduction to the Bootstrap (London: Chapman and Hall)
[20] Fisher R A 1915 Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population *Biometrika* **10** 507–21
[21] Fisher R A 1921 On the ‘probable error’ of a coefficient of correlation deduced from a small sample *Metron* **1** 3–32
[22] Gläser M, Borys M, Ratschko D and Schwartz R 2010 Redefinition of the kilogram and the impact on its future dissemination *Metrologia* **47** 419–28
[23] Haddad D, Seifeit F, Chao L, Possolo A, Newell D B, Pratt J R, Williams C J and Schlamminger S 2017 Measurement of the Planck constant at the National Institute of Standards and Technology from 2015 to 2017 *Metrologia* **54** 633–41
[24] Higgins J P T and Thompson S G 2002 Quantifying heterogeneity in a meta-analysis *Stat. Med.* **21** 1539–58
[25] Higham N J 2002 Computing the nearest correlation matrix—a problem from finance *IMA J. Numer. Anal.* **22** 329–43
[26] Hoaglin D C 2016 Misunderstandings about $Q$ and ‘Cochran’s $Q$ test’ in meta-analysis *Stat. Med.* **35** 485–95
[27] International Organization of Legal Metrology Weights of classes $E_1$, $E_2$, $F_1$, $F_2$, $M_{2/3}$, $M_2$, $M_{2/3}$, and $M_3$—Part 1: Metrological and technical requirements. *Bureau International de Métrologie Légale (OIML)*, Paris, France, 2004. https://www.oiml.org/en/files/pdf_rtr111-1-e04.pdf, International Recommendation OIML R 111-1 Edition 2004 (E)
[28] Koepke A, Lafarge T, Possolo A and Toman B 2017 Consensus building for interlaboratory studies, key comparisons, and meta-analysis *Metrologia* **54** S34–62
[29] Koepke A, Lafarge T, Toman B and Possolo A 2017 NIST Consensus Builder—User’s Manual (Gaithersburg, MD: National Institute of Standards and Technology)
[30] Kuramoto N, Mizushima S, Zhang L, Fujita K, Azuma Y, Kurokawa A, Okubo S, Inaba H and Fujii K 2017 Determination of the Avogadro constant by the XRCd method using a $^{28}$Si-enriched sphere *Metrologia* **54** 716
[31] Li Z, *et al.* 2017 The first determination of the Planck constant with the joule balance NIM-2 *Metrologia* **54** 763
[32] Lin L and Chu H 2016 Altmeta: alternative meta-analysis methods https://CRAN.R-project.org/package=altmeta R package version 2.2
[33] Lin L, Chu H and Hodges J S 2017 Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies *Biometrics* **73** 156–66
[34] Mohr P J and Taylor B N 2000 CODATA recommended values of the fundamental physical constants: 1998 *Rev. Mod. Phys.* **72** 351–495
[35] Mohr P J and Taylor B N 2005 CODATA recommended values of the fundamental physical constants: 2002 *Rev. Mod. Phys.* **77** 1–107
[36] Mohr P J, Taylor B N and Newell D B 2008 CODATA recommended values of the fundamental physical constants: *Rev. Mod. Phys.* **80** 633–730
[37] Mohr P J, Taylor B N and Newell D B 2012 CODATA recommended values of the fundamental physical constants: *Rev. Mod. Phys.* **84** 1527–605
[38] Mohr P J, Newell D B and Taylor B N 2016 CODATA recommended values of the fundamental physical constants: *Rev. Mod. Phys.* **88** 035009
[39] Newell D B *et al.* 2018 The CODATA 2017 values of $h$, $e$, $k$, and $N_A$ for the revision of the SI *Metrologia* accepted (https://doi.org/10.1088/1681-7575/aa950a)
[40] Quinn T 2012 From Artefacts to Atoms: the BIPM and the Search for Ultimate Measurement Standards (New York: Oxford University Press)
[41] R Core Team 2017 R: A Language and Environment for Statistical Computing (Vienna: R Foundation for Statistical Computing) (www.R-project.org/)
[42] Stock M, Barat P, Davis R S, Picard A and Milton M J T 2015 Calibration campaign against the international prototype of the kilogram in anticipation of the redefinition of the kilogram part I: comparison of the international prototype with its official copies *Metrologia* **52** 310–6
[43] Stock M *et al.* 2017 Report on CCM Pilot Study CCM.R-kg-P1: comparison of future realizations of the kilogram—final report *Technical Report* Bureau International des Poids et Mesures (BIPM), Sèvres, France
[44] Taylor B N, Parker W H and Langenberg D N 1969 Determination of $\rho_0$ using macroscopic quantum phase coherence in superconductors: implications for quantum electrodynamics and the fundamental physical constants *Rev. Mod. Phys.* **41** 375–496
[45] Thomas M, Ziane D, Pinot P, Karcher R, Imanaliev A, PereiraDos Santos F, Merlet S, Piquemal F and Espel P 2017 A determination of the Planck constant using the LNE Kibble balance in air *Metrologia* **54** 468–80
[46] Thompson M and Ellison S L R 2011 Dark uncertainty *Accreditation Qual. Assur.* **16** 483–7
[47] Viechtbauer W 2010 Conducting meta-analyses in R with the metafor package *J. Stat. Softw.* **36** 1–48
[48] Wood B M, Sanchez C A, Green R G and Liard J O 2017 A summary of the Planck constant determinations using the NRC Kibble balance *Metrologia* **54** 399–409
[49] Wood S N 2006 *Generalized Additive Models: an Introduction with R* (Boca Raton, FL: Chapman and Hall)