Consensus Value for Big G

Antonio Possolo and Blaza Toman

February 10, 2016

ABSTRACT

The dispersion of the estimates of G that have been considered in multiple CODATA adjustments of the fundamental constants, reflects a persistent challenge in the measurement of small forces. While this problem in precision measurement remains open, the statistical problem of combining the measured values into a recommended value needs to be addressed every time an adjustment of the fundamental constants is done.

The value that the CODATA Task Group on Fundamental Constants recommends for G in the 2014 adjustment is $6.674\,08 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, with associated standard uncertainty $u(G) = 0.000\,31 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$.

An alternative evaluation starts from a laboratory random effects model, which may be fitted to the data using any one of several different procedures. The maximum likelihood procedure, which can take into account correlations between estimates produced in different experiments or laboratories, produces the consensus estimate $6.673\,81 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, with associated standard uncertainty $u(G) = 0.000\,31 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$.

Both the maximum likelihood results, and the results of Bayesian analyses that we also describe, are statistically indistinguishable from one another and from the 2014 CODATA recommended value for G, yet none involves any potentially controversial inflation of the original, laboratory-specific uncertainty evaluations, and all were produced using established, widely used statistical methods.

In addition, we also quantify the extent of the disagreement between the values measured by the different laboratories, above and beyond the differences that would be expected based only on the stated laboratory-specific standard uncertainties: the standard deviation gauging that disagreement is about five times larger than the median of these standard uncertainties.

We have also explored the idea that when different measurement methods produce estimates of G in close mutual agreement, then such consilience may suggest greater proximity to the true value of the measurand than a comparably close agreement between results obtained by application of variants of the same measurement method. We propose a procedure to implement this idea based on relevant characteristics of subsets of the original data: the result, however, also is statistically indistinguishable both from the 2014 CODATA recommended value and from the maximum likelihood and Bayesian results.

1 Introduction

During the workshop held at NIST in October 9–10, 2014, on the Newtonian constant of gravitation, we learned that the central problem with the dispersion of the values that have been obtained for G in independent experiments, is a problem in precision measurement, not a statistical problem.

However, the periodic CODATA adjustment of the fundamental constants involves solving the statistical problem of how best to combine the available measurement results into a consensus value, and how to qualify this value with an evaluation of uncertainty.

The purpose of this note is to describe how this may be done using a statistical model and widely used statistical data reduction procedures, as alternative to how CODATA has addressed the problem: an alternative that is well-aligned with how similar problems have been addressed in other disciplines, in particular in *meta-analysis* in medicine.

First, in a preamble (§2), we review several issues that we had the pleasure of discussing with Jim Faller, Barry Taylor, and Stephan Schlamminger on Friday, February 5th, 2016, in the library of the Statistical Engineering Division, at NIST, in Gaithersburg, MD.

Second, in §3, we describe the laboratory effects model that would be the choice that many statisticians would make when confronted with the task of combining the measurement results available for G into a single estimate of its true value qualified with a statement of associated uncertainty.

Third, we present the results of fitting the laboratory effects model to the data used by CODATA to produce the 2014 recommended value for G, in two different ways: using the method of maximum likelihood estimation (MLE, in §4), and Bayesian methods (with several variants, described in §5). The results are very similar to one another, except that the Bayesian methods produce somewhat larger uncertainties than the MLE's, and one of them, the linear pool (§5.4) produces an uncertainty strikingly larger than all the others — this may also be the uncertainty that some scientists may find is the more believable.

Since the results of these alternative approaches are very similar also to the 2014 CODATA results, the obvious conclusion is that there would be no downsides, and only upsides, were CODATA to choose to adopt a tried and trusted method to combine measurement results for G resting on a statistical model that is widely recognized as being generally well suited to this type of data reduction.

Finally, in §6, we sketch a new approach to the calculation of the consensus value exploiting a suggestion of David Newell's, whereby the close mutual agreement of a subset of values of G that will have been measured by fundamentally different methods, should be taken not merely as a fortuitous coincidence, but instead as possibly indicative of proximity to the true value of G, and weighed more heavily when results are combined.

2 Preamble

A particularly striking feature of Exhibit 2 is the small margin of uncertainty surrounding the different estimates of the consensus value for G, by comparison with the rather large dispersion of the measured values that the consensus value summarizes. How can this possibly be?

First we point out that the consensus value, however it may be arrived at, is some sort of (weighted) average of the measured values. If the $n = 14$ values listed in the column headed "G" in Exhibit 1 are regarded as a sample from the a probability distribution (that is, like outcomes of independent random variables with the same probability distribution), then the standard uncertainty associated with their average is $s/\sqrt{14} = 0.000\,31 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$, which happens to be the value that CODATA currently assigns to $u(G)$.

This evaluation of standard uncertainty for the simple average of the measured values is just a trifle larger (1.16 times larger, to be precise) than the median of the standard uncertainties associated with the 14 values being averaged. However, that same standard uncertainty is 3.7 times *smaller* than the standard deviation of the values being averaged.

This is a consequence of the law of averages, which says that, as the sample size increases, the standard uncertainty of the sample average decreases in proportion to the square root of the sample size ($\sqrt{14} = 3.7$). However, this still seems to fall short of allaying the uneasiness that many people feel when they are told that, on the one hand, $u(G) = 0.000\,31 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$, while, on the other hand, they can see that the dispersion of values in Exhibit 2 is quite something else.

We believe that the reason for this uneasiness is that $u(G)$ is the answer to a question different from the question that people have in mind who find its value too small by comparison with what Exhibit 2 shows. The different question is this: if one were to make an independent measurement of G using one of the methods represented in this data set, with comparable experimental

acumen, where would the resulting value likely lie?

Or, in other words, what is the standard uncertainty of a prediction for a comparable, future measurement, based on these measured values that we have in hand? The answer is $0.001\,26 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$, which is almost 10 times larger than $u(G)$: the corresponding interval is depicted by a thin, dotted (gray) line in Exhibit 2. The assumption that validates this answer is that the 14 measured values are like a sample from a Gaussian probability distribution. It just so happens that Anderson and Darling (1952)'s statistical test of normality (that is, of Gaussian shape) finds no reason to dismiss such assumption.

None of the foregoing considerations and results utilize the uncertainties quoted by the different laboratories, and that are listed in the column headed "$u(G)$" in Exhibit 1. In the next sections we will reanalyze the data taking these uncertainties into account. The final conclusions, however, will be hardly different from the conclusions of the naive analysis we have just presented.

This possibly surprising outcome is a consequence of the fact that the individual uncertainties associated with the measured values are so much smaller than the actual dispersion of the measured values, which becomes apparent only once these values are inter-compared.

Such dismal state of affairs may truly, if allegorically, be attributed to the existence of yet unidentified sources of *dark uncertainty* (Thompson and Ellison, 2011), and motivates the rather unnerving observation that "while the measurement accuracy of little g has increased by eight orders of magnitude during its 400-year measurement history, the measurement accuracy of big G has only increased by three orders of magnitude during its 300-year measurement history" (Faller, 2014).

3 Laboratory Effects Model

Exhibit 1 lists the data used by CODATA to estimate G and to evaluate the associated uncertainty. An alternative reduction of these data may be undertaken using methods that have been widely used to combine the results of multiple studies, in particular in medicine, where such combination is known as *meta-analysis* (Cooper et al., 2009).

The basis for this alternative data reduction is a statistical model for the measurement results. Supposing that n laboratories made measurements, the measurement result from laboratory j comprises an estimate x_j of G and an

evaluation u_j of the associated standard uncertainty, for $j = 1, \ldots, n$.

The statistical model expresses the value measured by laboratory j as $x_j = G + \lambda_j + \epsilon_j$, where λ_j denotes the laboratory effect, and ϵ_j denotes measurement error. Both the laboratory effects $\lambda_1, \ldots, \lambda_n$ and the measurement errors $\epsilon_1, \ldots, \epsilon_n$ are modeled as outcomes of random variables with mean zero. The $\{\lambda_j\}$ all have the same standard deviation τ, but the $\{\epsilon_j\}$ may have different standard deviations $\{u_j\}$. Toman and Possolo (2009, 2010) discuss the use of laboratory effects models in measurement science, and Higgins et al. (2009) review them in general.

It is the presence of the $\{\lambda_j\}$ that gives this model its name, laboratory *random effects* model, and that allows it to accommodate situations where the variability of the estimates $\{x_j\}$ exceeds what would be reasonable to expect in light of the associated uncertainties $\{u_j\}$. As is obvious from Exhibit 2 such excess variability is a major source of uncertainty in this case.

All of these random variables usually are assumed to be Gaussian and independent, but neither is necessarily the case. In fact, when this model is used to estimate the value of G in the context of the CODATA adjustment, correlations between some of the laboratory effects need to be taken into account. In some applications, either the laboratory effects, or the measurement errors, or both, may have non-Gaussian distributions (Rukhin and Possolo, 2011).

4 Maximum Likelihood Estimation

The model may be fitted to the data listed in Exhibit 1 using any one of several different statistical procedures. For example, DerSimonian and Laird (1986) introduced one of the more widely used procedures, and Toman (2007) and Bodnar et al. (2016b) describe Bayesian procedures.

The more popular procedures assume that the laboratory effects and the errors are mutually independent. Since, in this case, the laboratory effects for NIST-82 and LANL-97 are correlated with correlation coefficient 0.351 (Mohr et al., 2012, Pages 1568–1569), and the laboratory effects for HUST-05 and HUST-07 are correlated with correlation coefficient 0.134 (D. Newell, 2015, personal communication), the more popular fitting procedures are not applicable here.

The method of maximum likelihood estimation may be used to fit the laboratory effects model to the data even in the presence of such correlations.

	$/1 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$	
	G	$u(G)$
NIST82	6.672 482	0.000 428
TRD96	6.672 900	0.000 500
LANL97	6.673 984	0.000 695
UWash00	6.674 255	0.000 092
BIPM01	6.675 590	0.000 270
UWup02	6.674 220	0.000 980
MSL03	6.673 870	0.000 270
HUST05	6.672 220	0.000 870
UZur06	6.674 252	0.000 124
HUST09	6.673 490	0.000 180
JILA10	6.672 340	0.000 140
BIPM13	6.675 540	0.000 160
ROSI14	6.671 910	0.000 990
UCI14	6.674 350	0.000 126

Exhibit 1: Values of G and $u(G)$ used to determine the 2014 CODATA recommended value, together with correlations of 0.351 (between NIST82 and LANL97), and of 0.134 (between HUST05 and HUST09).

(Bayesian methods can do the same.) The general idea of maximum likelihood estimation is to choose values for the quantities whose values are unknown (G and τ in this case) that maximize the probability (density) of the data. Application of this method requires that the probability distribution of the random variables that the data are conceived as realized values of, be modeled explicitly.

We assume that the joint probability distribution of the $\{x_j\}$ is multivariate Gaussian with n-dimensional mean vector all of whose entries are equal to G (meaning that all the laboratories indeed are measuring the same quantity), and with covariance matrix $S = U + V$. Both U and V are $n \times n$ symmetric matrices. The entries of the main diagonal of U are all equal to τ^2, and the off-diagonal entries are all zero, except those that correspond to the pairs of laboratories mentioned above: $0.351\tau^2$ or $0.234\tau^2$, depending on the pair. V denotes a diagonal matrix with the $\{u_j^2\}$ in the main diagonal.

The probability density to be maximized with respect to G and τ is $f(\boldsymbol{x}|G,\tau) = (2\pi)^{-n/2}|S^{-1}|^{1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$, where $\top$ denotes matrix transposition, $\boldsymbol{x} = (x_1,\ldots,x_n)^\top$ $\boldsymbol{\mu} = (G,\ldots,G)^\top$ are column vectors, and S is as defined above, with inverse S^{-1}, and $|S^{-1}|$ denotes the determinant of its inverse.

The maximization was done numerically, under the constraints that both G and τ be non-negative, using function `nloptr` defined in the package of the same name for the R environment for statistical computing and graphics (Ypma, 2014; Johnson, 2015; R Core Team, 2015), using the "Subplex" algorithm (Rowan, 1990).

According to the theory of maximum likelihood estimation (Wasserman, 2004), the results of the optimization can also be used to obtain an approximation for $u(G)$. The quality of this approximation generally tends to increase with increasing number n of laboratories. We have employed the parametric statistical bootstrap (Efron and Tibshirani, 1993) to validate this approximation.

Based on the data in Exhibit 1, and with the modeling assumptions just described, the maximum likelihood estimate of G is $6.673\,80 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, with approximate associated standard uncertainty $u(G) = 0.000\,30 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$. This consensus value and standard uncertainty are depicted in Exhibit 2. The same results are obtained when the correlations aforementioned are neglected. Obviously, the maximum likelihood estimate and the 2014 CODATA recommended value are statistically indistinguishable once the corresponding associated uncertainties are taken

into account.

The standard deviation τ, of the laboratory effects $\lambda_1, \ldots, \lambda_n$, also is of scientific interest because it quantifies the extent of the disagreement between the values measured by the different laboratories, above and beyond the differences that would be expected based only on the stated laboratory-specific standard uncertainties $\{u_j\}$.

The maximum likelihood estimate of τ is $0.001\,02 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$, which is 3.8 times larger than the median of the $\{u_j\}$, suggesting that there may be very substantial sources of uncertainty still to be characterized that are responsible for that disagreement.

One shortcoming of the uncertainty evaluation described above is that it fails to take into account the uncertainty surrounding the estimate of the between-laboratories standard deviation τ, when in this case this estimate is based on a fairly small number of degrees of freedom (we have only 14 measured values to gauge it from). Bayesian methods address this shortcoming, and we turn to them next.

5 Bayesian Estimation

The statistical model used for Bayesian estimation is the same laboratory effects model introduced above, which expresses the value measured by laboratory j as $x_j = G + \lambda_j + \epsilon_j$, with the same assumptions for the $\{\epsilon_j\}$ that were already made. Unrealistic as it may be, we will continue to treat the $\{u_j\}$ as if they were based on infinitely many numbers of degrees of freedom.

The Bayesian paradigm for statistical analysis is both very simple to state and very useful in applications. It amounts to treating all quantities whose values are unknown as values of non-observable random variables, and the experimental data as observed outcomes of other random variables. Therefore, we need to assign probability distributions to G itself, to the $\{\lambda_j\}$, and also to τ and ν, regarding them as values of non-observable random variables.

5.1 Conventional Random Effects Model

This approach offers the flexibility and the technical means to incorporate coherently any relevant information about the quantities (G in particular) whose true values are sought, and that may be in hand prior to obtaining the fresh experimental data.

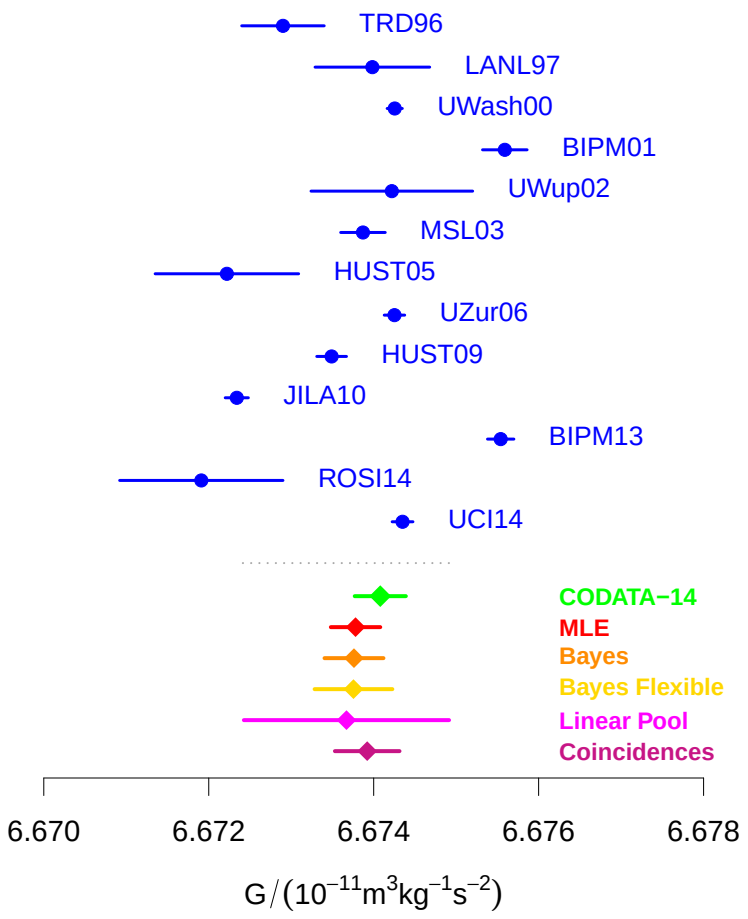

Exhibit 2: Measurement results for G used in the CODATA 2014 adjustment, the corresponding recommended value of G and associated standard uncertainty, and their counterparts obtained by application of the method of maximum likelihood, the Bayesian procedure described in §5, and the method based on weighted subsets described in §6. The measurement results are depicted in blue, with a dot indicating the measured value, and the horizontal line segment representing the interval $x_j \pm u_j$. The thin (gray) dashed line represents a 68 % coverage interval for a *prediction* of a future measured value, made comparably to those already represented in the data set, as discussed in §2. The 2014 CODATA recommended value and associated standard uncertainty, and their counterparts corresponding to maximum likelihood (§4) and Bayesian estimation (§5), as well as for the data reduction that takes into account mutual agreement and methodological diversity of subsets of the measurements (§6), are depicted similarly. Obviously, the 2014 CODATA recommended value and the five alternative results are statistically indistinguishable.

In our case, and since we are using measurement results obtained since 1982, it could be argued that we should use as prior information about G the knowledge about its value that had accumulated until that date (Speake and Quinn, 2014).

We will not attempt such exercise here, and will introduce only the belief that we know the order of magnitude of G by assigning to it a uniform (or, rectangular) distribution concentrated on the interval from $1 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ to $10 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$.

Differently from what was assumed in §4, we will model the $\{\lambda_j\}$ as a sample from a re-scaled Student t distribution with unknown scale τ and number ν of degrees of freedom, because this provides flexibility to accommodate measured values that may lie far from the bulk of the others (Kruschke, 2013; Possolo, 2013).

We model the scale parameter τ as a random variable with a half-Cauchy distribution, with scale $15 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, hence highly non-informative, following a recommendation made by Gelman (2006).

The prior distribution assigned to ν, the number of degrees of freedom of the Student t distribution assumed for the $\{\lambda_j\}$, is the distribution of $Y + 1$, where Y has an exponential distribution with mean 29. This is the same choice made by Kruschke and Meredith (2013) for R package `BEST`.

Given the values of G, the $\{\lambda_j\}$, and τ, we do otherwise continue to model the measured values of G as outcomes of Gaussian random variables just the same as in §4, except that here we disregard the correlations between two pairs of laboratories because they are inconsequential.

Having thus modeled both unknowns (G and τ) and the data, pursuing the Bayesian approach (Gelman et al., 2013) next involves application of Bayes' rule to compute the conditional probability distribution (*posterior* distribution) of G, the $\{\lambda_j\}$, ν, and τ, given the data.

For the modeling choices we made, this distribution cannot be computed in closed form, and we have had to resort to Markov Chain Monte Carlo (MCMC) sampling instead, to draw a large sample from such distribution, wherefrom we have derived estimates and uncertainties for G and τ. We did this using JAGS (Plummer, 2015) with the facilities implemented in R package `R2jags` (Su and Yajima, 2015; R Core Team, 2015).

We have verified that the MCMC sampling process reached equilibrium by running three chains concurrently and applying the convergence diagnostic test suggested by Geweke (1998) to all of the unknown quantities, taking into account the multiplicity of simultaneous testing by adjusting the resulting

p-values using the procedure suggested by Benjamini and Hochberg (1995).

The corresponding estimate of G is $6.673\,76 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, with associated standard uncertainty $u(G) = 0.000\,36 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$. These results are depicted in Exhibit 2. The estimate of τ is $0.001\,35 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, which is 5 times larger than the median of the $\{u_j\}$. The mean of the posterior distribution of ν was 32 (with standard deviation 30), indicating that the $\{\lambda_j\}$ are effectively like a sample from a Gaussian distribution, yet suggesting that this conclusion is surrounded by considerable uncertainty.

An alternative modeling choice, using the *reference* prior distribution for G and τ described by Bodnar et al. (2016a), produces $6.673\,80 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ as estimate of G, $0.000\,29 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ for the associated standard uncertainty, and $0.000\,94 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ as estimate of τ.

5.2 Flexible Random Effects Model

Because the assumption used in the previous subsection, that the laboratory effects $\{\lambda_j\}$ are a sample from a single Student's t distribution may be considered too restrictive, especially when the measured values are very dispersed, it is useful to relax that assumption and entertain models where different subsets of the $\{\lambda_j\}$ are samples from different distributions.

There are various schemes for selecting subsets (or, *clusters*) of identically distributed $\{\lambda_j\}$. Most of these schemes are based on a Dirichlet process, which is a probability distribution on a class of probability distributions (Escobar and West, 1995). The $\{\lambda_j\}$ are regarded as a sample from a common distribution P about which we are uncertain. We model this uncertainty by representing P as a mixture of Gaussian distributions whose mixing probabilities are drawn from a Dirichlet distribution.

Applying the modeling and computational technique described by Ohlssen et al. (2007) with $N = 28$ (the number of components in the mixture mentioned above), using MCMC sampling with three chains, whose equilibrium we verified, and also having assessed model adequacy, we obtained $6.673\,75 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ as estimate of G, with associated standard uncertainty $u(G) = 0.000\,47 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, depicted in Exhibit 2.

The main clusters of laboratories found in the process were {`LANL97`, `UWash00`, `UZur06`, `UCI14`}, {`NIST82`, `JILA10`, `ROSI14`}, and {`BIPM01`, `BIPM13`}.

m	K	$/1\times10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ G	$u(G)$
1	14	6.673 67	0.001 24
2	91	6.673 86	0.000 89
3	364	6.673 99	0.000 63
4	1001	6.674 12	0.000 45

Exhibit 3: Bayesian model averaging estimates of G, and associated standard uncertainties, corresponding to subsets of measured values of sizes $m = 1, 2, 3, 4$. K denotes the number of subsets corresponding to each value of m.

5.3 Bayesian Model Averaging

Bayesian model averaging (Hoeting et al., 1999) serves to combine estimates produced by alternative statistical models based on posterior probabilities of the models given the data. Therefore, it affords the means to entertain all alternative models simultaneously without having to adopt any one of them once and for all.

In our case the alternative models are defined as follows (Elster and Toman, 2010): consider a subset of size m of the 14 measured values of G, and suppose that the values in this set are without bias, that is, the mean of each one of them is G; for the remaining $14-m$ not in this subset we suppose that their means deviate from G by the corresponding laboratory effects $\{\lambda_j\}$ — that is, their means are of the form $G+\lambda_j$. We will use a non-informative prior distribution on G and on the $14-m$ of the $\{\lambda_j\}$.

For example, if $m = 4$ then there are 1001 different subsets of 4 measured values out of 14. Each of these subsets is used to produce an estimate of G and a posterior probability of the particular subset satisfying the model assumption (that all 4 values in the subset are unbiased estimates of G). These posterior probabilities are then used as weights to combine the 1001 estimates of G into a weighted average, and to produce a corresponding evaluation of standard uncertainty. Exhibit 3 lists the results for subset sizes $m = 1, 2, 3, 14$.

The case of $m = 1$ is special and particularly interesting because it is the same as the so-called *linear pool* method for reconciling expert opinions expressed in the form of probability distributions, which we review next.

5.4 Linear Pool

The linear pool was suggested by Stone (1961) but Bacharach (1979) attributes the idea to Pierre Simon, Marquis de Laplace (Genest and Zidek, 1986).

The "experts" in our case are the 14 different laboratories that measured G, and the probability distributions that encapsulate their "opinions" are taken as Gaussian probability distributions with means equal to the values they measured for G, and standard deviations equal to the corresponding standard uncertainties, listed in Exhibit 1.

Application of the linear pool is equivalent to asking each "expert" to cast a vote in favor of her work, where her vote is a random draw from the probability distribution that encapsulates the expert's assessment and that expresses her confidence in her results. The final result is a probability distribution that represents the collective state of knowledge of all the "experts" about the value of G.

We produced a sample from this distribution by repeating the following two steps 1×10^6 times: (i) select one laboratory uniformly at random from among the 14 that are under consideration (each one with the same probability of being selected); and (ii) draw one value from the selected laboratory's Gaussian distribution. Exhibit 4 shows a smooth histogram of this sample, and indicates its mean (the consensus estimate of G) and standard deviation (evaluation of $u(G)$), for which there are closed form solutions: the former is the arithmetic mean of the 14 measurements, and the latter is $\left(\sum_{j=1}^{n}[u^2(x_j)+(x_j-\overline{x})^2]/n\right)^{1/2}$ (Elster and Toman, 2010).

6 Meaningful Coincidences

D. Newell (2015, personal communication) suggested that the close proximity (*coincidence*) of measured values obtained by two or more different measurement methods is potentially significant from a substantive viewpoint: examples include the triplet {`UWash00`, `UZur06`, `UCI14`}, and the pair {`NIST82`, `JILA10`} (Exhibit 2).

Stimulated by this suggestion, we have developed a framework whereby a consensus value for G will obtain by combining estimates derived from subsets of the measured values, weighed according to both a *diversity index* and a *concentration index*.

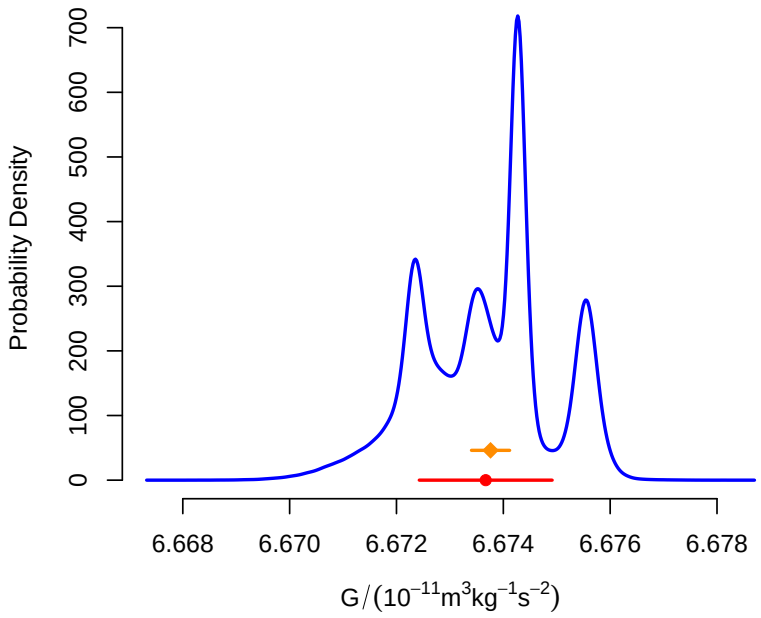

Exhibit 4: Smooth histogram of the sample drawn from the probability distribution that reconciles the measurement results for G by application of the linear pool. The (red) dot indicates the mean of the sample, $6.673\,67\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, and the superimposed (red) segment has length twice the standard deviation of the sample and is centered at the sample mean. The (orange) diamond and line segment depict the results of the Bayesian data reduction described in §5.1, which are also depicted in Exhibit 2. Therefore, the linear pool produces an estimate of G that is statistically indistinguishable from all the others, but a considerably larger standard uncertainty, $0.001\,24\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$.

The diversity index reflects how varied a subset is in terms of the measurement methods represented in it: the more varied, the larger the weight. The concentration index reflects how closely together the values lie that were measured by the laboratories in the subset: the more concentrated (that is, the less dispersed), the greater the weight.

This framework is inspired by the use of subsample values as typical values (Hartigan, 1969, 1975; Atkins and Sherman, 1992). The simplest version of the idea goes like this: suppose that $X_1, \ldots, X_n$ are independent random variables with continuous distributions symmetrical about θ, and let $A_1, A_2, \ldots, A_K$ denote the averages of all $K = 2^n - 1$ non-empty subsets of the $\{X_i\}$, whose ordered values are $A_{(1)} \leqslant A_{(2)} \leqslant \cdots \leqslant A_{(K)}$.

In these circumstances, each of the $K + 1$ intervals $(-\infty, A_{(1)})$, $(A_{(1)}, A_{(2)})$, …, $(A_{(n-1)}, A_{(K)})$, and $(A_{(K)}, +\infty)$, covers θ with probability $1/(K + 1)$, and the subset averages $\{A_k\}$ are said to be *typical values*. Hartigan (1969) proved that one need not consider all non-empty subsets, and that the averages for a suitably *balanced* sample drawn from the set of all subsets also are typical values, in the sense aforementioned.

The idea here is to define A_k as the consensus value corresponding to the kth subset of the measured values in Exhibit 1, and then combine the consensus values corresponding to all the non-empty subsets using suitable weights. In our case there are only $K = 2^{14} - 1 = 16\,383$ non-empty subsets of the measured values, and it is practicable to compute consensus values for all of them.

However, since the computer execution of the procedures described in §4 and in §5 is fairly time-consuming, for the present purpose we employ the procedure proposed by DerSimonian and Laird (1986) (DL), because it produces the results about instantaneously as implemented in R function `rma`, which is defined in package `metafor` (Viechtbauer, 2010). It also just so happens that this DL procedure is one of the most widely used procedures for meta-analysis.

Exhibit 5 shows a smooth histogram of the DL consensus values corresponding to the 16 383 non-empty subsets of laboratories, and a 68 % coverage interval (corresponding to plus or minus one standard uncertainty) defined as the union of 11 185 intervals between successive ordered consensus values centered around the value of G at which the smooth histogram peaks. This coverage interval is of just about the same length as the corresponding interval computed using the Bayesian procedure described in §5.1, which is also shown in the same Exhibit 5.

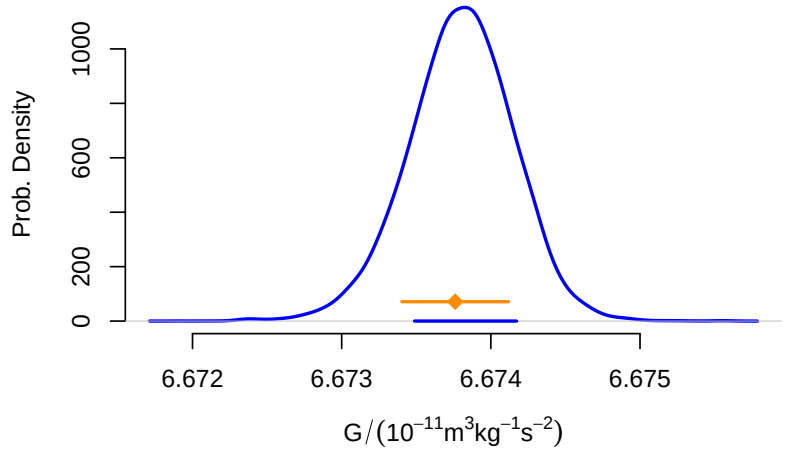

Exhibit 5: Smooth histogram of the consensus estimates of G derived from the 16 383 non-empty subsets of the 14 measured values listed in Exhibit 1, using the DL procedure (DerSimonian and Laird, 1986). Also shown are the Bayes estimate (orange diamond) adorned with a line segment whose length represents plus or minus one standard uncertainty, and below it a horizontal (blue) segment representing a 68 % coverage interval for the true value of G derived from the subset values.

Subset number 339 (in the particular order that R function `combn` generates all these subsets) is $S_{339} = \{\texttt{UWash00}, \texttt{UZur06}, \texttt{UCI14}\}$: the corresponding estimates of G and of τ are $6.674\,28 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$, and $0\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$. For subset $S_{3537} = \{\texttt{NIST82}, \texttt{TRD96}, \texttt{LANL97}, \texttt{BIPM01}, \texttt{HOUST05}, \texttt{ROSI14}\}$ the estimates of G and of τ are $6.673\,26 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ and $0.001\,71 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$.

In S_{339} the measurement methods are all different: fiber torsion (dynamic) for `UWash00`, stationary body (weight change) for `UZur06`, and cryogenic torsion pendulum for `UCI14`. In S_{3537}, fiber torsion (dynamic) occurs four times (for `NIST82`, `TRD96`, `LANL97`, and `HUST05`), strip torsion (compensation) occurs once (for `BIPM01`), and cooled atoms (interferometry) also occurs only once (for `ROSI14`).

We use the entropy of the relative frequencies with which the different measurement methods occur in a subset to define the value of the diversity index associated with the subset, and the reciprocal of the estimate of τ^2 for the laboratories in the subset, to define the corresponding value of the concentration index.

For example, for S_{339} the average entropy per laboratory is $H_{339} = -((1/3)\log(1/3) + (1/3)\log(1/3) + (1/3)\log(1/3))/3 = 0.366$, and for S_{3537} it is $H_{3537} = -((1/6)\log(1/6) + (4/6)\log(4/6) + (1/6)\log(1/6))/6 = 0.145$, therefore indicating that S_{339} is methodologically more diverse than S_{3537}. The corresponding weights are proportional to $\exp(H_{339})$ and $\exp(H_{3537})$.

The weights corresponding to the concentration index are proportional to the reciprocal of the sum of the estimate of τ^2 and the median of the squared laboratory-specific uncertainties $\{u_j^2\}$ for the laboratories represented in the subset. The overall weights are the product of the entropy and concentration weights, normalized to add up to 1.

We use these weights to define a weighted version of the empirical cumulative distribution function (CDF) for the $K = 16\,383$ DL consensus values. The empirical CDF of these values is an increasing step function that is 0 for all values smaller than the smallest of the subsample values, and increases by $1/K$ at each of the (ordered) consensus values (which happen to be all different from one another), reaching its maximum of 1 at the largest of the consensus values. We modify the empirical CDF by replacing the identical step increases of $1/K$ by step increases equal to the weights corresponding to the consensus values. Since these weights were previously normalized to sum to 1, this weighted empirical CDF is still a *bona fide* CDF, being non-negative and increasing monotonically from 0 to 1.

To compute the mean and the variance of the weighted set of consensus values, we use the following formula for the kth moment of a positive random variable with CDF F, for $k = 1, 2, \ldots$, which is a consequence of Tonelli's theorem (Tonelli, 1909; Royden, 1968, Chapter 12, Theorem 20): $k \int_0^\infty x^{k-1}(1 - F(x))\mathrm{d}x$.
The mean of the probability distribution corresponding to the weighted empirical CDF described above is the estimate of G produced by this framework that exploits coincidences: $6.673\,92 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$. The associated standard uncertainty is $u(G) = 0.000\,39 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$ (Exhibit 2).

7 Conclusions

In this study we have applied six different statistical methods of data reduction to the values measured for G by 14 different laboratories and that were used by CODATA to produce their 2014 recommended value for G, and associated uncertainty. All of these statistical methods correspond to alternative sets of assumptions for the same laboratory random effects model that expresses each measured value as $x_j = G + \lambda_j + \epsilon_j$ for $j = 1, \ldots, n = 14$:

(1) Maximum likelihood estimation under a Gaussian model, considering stated correlations between two pairs of measured values (§4);

(2) Conventional Bayesian estimation using MCMC and half-Cauchy prior distributions for the between-laboratory, and within-laboratory uncertainty components (§5.1);

(3) Flexible random effects model fitted to the data using a Bayesian procedure involving MCMC sampling (§5.2);

(4) Bayesian model averaging, also involving MCMC sampling (§5.3);

(5) Linear pool using a very simple, conventional Monte Carlo procedure (§5.4); and

(6) Meaningful coincidences between measurement methods using a suitably weighted empirical cumulative distribution function of typical values (§6).

The alternative results all are statistically indistinguishable from one another, and also from the 2014 CODATA recommended value, once their respective

uncertainties are taken into account. Since several of the methods illustrated in this study (in particular those numbered (1), (2), and (5) above) have a very long history of usage, in particular for reductions of data obtained in inter-laboratory studies and in meta-analyses, we believe that adopting any one of these procedures will facilitate widespread acceptance of the consensus estimate of G. Furthermore, all of these statistical methods produce evaluations of uncertainty that can be as concise as the standard uncertainty, or as comprehensive as an arbitrarily large sample drawn from the probability distribution that describes the uncertainty surrounding the consensus value.

We have also explained (in §2) the apparently counter-intuitive fact that the standard uncertainties produced by all the methods we have applied (except the linear pool), as well as the standard uncertainty announced by CODATA, appear much too small when compared with the actual dispersion of the measured values, painfully obvious in Exhibit 2.

The explanation we offer hinges on the distinction between a coverage interval for consensus values that are very much like averages, and a coverage interval for a prediction of a future measurement of quality (and uncertainty) comparable to those available currently.

The fact that the linear pool produces a standard uncertainty that is comparable to that prediction uncertainty, hence much larger than all the other methods, suggests that this method of data reduction is producing a consensus value that indeed pools all the individual measurement results, but assigns to it an uncertainty that is essentially different from the conventional averaging methods.

Finally, while some might feel disheartened by the fact that the method that takes methodological coincidences into account produces results essentially in agreement with the other results, we choose to regard it in a positive light and take it as good news that multiple, very different methods of data reduction all seem to point very much towards the same, fairly narrow range of credible values for G.

References

T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

J. E. Atkins and G. J. Sherman. Sets of typical subsamples. *Statistics & Probability Letters*, 14(2):115–117, 1992. doi: 10.1016/0167-7152(92)90074-F.

M. Bacharach. Normal Bayesian dialogues. *Journal of the American Statistical Association*, 74(368):837–846, December 1979.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57:289–300, 1995.

O. Bodnar, C. Elster, J. Fischer, A. Possolo, and B. Toman. Evaluation of uncertainty in the adjustment of fundamental constants. *Metrologia*, 53(1): S46–S54, 2016a.

O. Bodnar, A. Link, and C. Elster. Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Analysis*, 11(1): 25–45, March 2016b. doi: 10.1214/14-BA933.

H. Cooper, L. V. Hedges, and J. C. Valentine, editors. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation Publications, New York, NY, 2nd edition, 2009.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, September 1986.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993.

C. Elster and B. Toman. Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using bayesian model averaging. *Metrologia*, 47:113–119, 2010.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

J. E. Faller. Precision measurement, scientific personalities and error budgets: the *sine quibus non* for big *g* determinations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372 (2026), 2014. doi: 10.1098/rsta.2014.0023.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall / CRC, Boca Raton, FL, 3rd edition, 2013.

C. Genest and J. V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, February 1986.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. Clarendon Press, Oxford, UK, 1998.

J. A. Hartigan. Using subsample values as typical values. *Journal of the American Statistical Association*, 64(328):1303–1317, December 1969. doi: 10.1080/01621459.1969.10501057.

J. A. Hartigan. Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *The Annals of Statistics*, 3 (3):573–580, May 1975. doi: 10.1214/aos/1176343123.

J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 172(1):137–159, January 2009.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

S. G. Johnson. The NLopt nonlinear-optimization package. `http://ab-initio.mit.edu/nlopt`, 2015. Last visited August 21, 2015.

J. K. Kruschke. Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, 142(2):573–603, 2013.

J. K. Kruschke and M. Meredith. *BEST: Bayesian Estimation Supersedes the t-Test*, 2013. URL `http://CRAN.R-project.org/package=BEST`. R package version 0.2.0.

P. J. Mohr, B. N. Taylor, and D. B. Newell. Codata recommended values of the fundamental physical constants: 2010. *Reviews of Modern Physics*, 84(4): 1527–1605, October-December 2012.

D. I. Ohlssen, L. D. Sharples, and D. J. Spiegelhalter. Flexible random-effects models using bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine*, 26(9):2088–2112, 2007. doi: 10.1002/sim.2666.

M. Plummer. *JAGS Version 4.0.0 user manual*, October 2015. URL `http://mcmc-jags.sourceforge.net/`.

A. Possolo. Five examples of assessment and expression of measurement uncertainty. *Applied Stochastic Models in Business and Industry*, 29:1–18, January/February 2013. doi: 10.1002/asmb.1947. Discussion and Rejoinder pp. 19–30.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

T. Rowan. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, University of Texas at Austin, Austin, TX, 1990. Department of Computer Sciences.

H. L. Royden. *Real Analysis*. MacMIllan Publishing Co., New York, NY, second edition, 1968.

A. L. Rukhin and A. Possolo. Laplace random effects models for interlaboratory studies. *Computational Statistics and Data Analysis*, 55:1815–1827, 2011.

C. Speake and T. Quinn. The search for newtonâĂŹs constant. *Physics Today*, 67(7):27–33, 2014. doi: 10.1063/PT.3.2447.

M. Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32: 1339–1342, December 1961.

Y.-S. Su and M. Yajima. *R2jags: Using R to Run 'JAGS'*, 2015. URL `https://CRAN.R-project.org/package=R2jags`. R package version 0.5-7.

M. Thompson and S. L. R. Ellison. Dark uncertainty. *Accreditation and Quality Assurance*, 16:483–487, 2011.

B. Toman. Bayesian approaches to calculating a reference value in key comparison experiments. *Technometrics*, 49(1):81–87, February 2007.

B. Toman and A. Possolo. Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 14:553–563, 2009.

B. Toman and A. Possolo. Erratum to: Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 15: 653–654, 2010.

L. Tonelli. Sull'integrazione per parti. *Atti della Accademia Nazionale dei Lincei (5)*, 18(2):246–253, 1909.

W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010. URL `http://www.jstatsoft.org/v36/i03/`.

L. Wasserman. *All of Statistics, A Concise Course in Statistical Inference*. Springer Science+Business Media, New York, NY, 2004.

J. Ypma. Introduction to `nloptr`: an R interface to NLopt, August 2014. URL `http://cran.fhcrc.org/web/packages/nloptr/`. Vignette for R package `nloptr`.