Alternative analyses of measurements of the Planck constant



2012 Metrologia 49 567

(http://iopscience.iop.org/0026-1394/49/4/567)

Metrologia **49** (2012) 567–571 doi:10.1088/0026-1394/49/4/567

# Alternative analyses of measurements of the Planck constant

**Blaza Toman**[1], **Joachim Fischer**[2] **and Clemens Elster**[2]

[1] Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, US Department of Commerce, Gaithersburg, MD, USA
[2] Physikalisch-Technische Bundesanstalt, Abbestrasse 2-12, 10587 Berlin, Germany

Received 9 April 2012, in final form 5 June 2012
Published 5 July 2012
Online at stacks.iop.org/Met/49/567

**Abstract**
We compare different methods for the adjustment of the Planck constant using recent, inconsistent data. First we discuss the popular Birge ratio method. By stating the underlying statistical model we critically assess and modify this method. We then propose alternative ways of adjusting the inconsistent data using different, less restrictive assumptions about the cause of the inconsistency. Finally, we discuss the results and give some guidance as to the choice of method.

(Some figures may appear in colour only in the online journal)

## 1. Introduction

The new definition of the kilogram will be based on the Planck constant $h$ [1]. According to the present definition, the uncertainty of the mass of the international kilogram prototype is zero by convention. However, any new realization will allocate an uncertainty to the kilogram. Two experiments have contributed significantly to the CODATA value of $h$ of 2010 and have, together with the reestablished NPL watt balance experiment at NRC in Canada, the potential to achieve the agreed goal for the relative uncertainty of $2 \times 10^{-8}$. One is the NIST watt balance experiment which aims at measuring $h$ by the virtual comparison of mechanical and electrical powers [2]. The other experiment aims at determining the Avogadro constant $N_A$ by counting the atoms in a nearly perfect single-crystal silicon sphere highly enriched with the $^{28}$Si isotope [3]. A relative $1.8 \times 10^{-7}$ discrepancy has been observed when comparing the results of these experiments through the molar Planck constant, $N_A h$. This discrepancy is not consistent with the stated uncertainties and therefore needs further consideration. This paper presents several alternative statistical models to combine the available data and form a consensus estimate of $h$. All analyses resolve the inconsistency of the measurements. Recommendations as to which model may be preferred are given in section 5.

The starting point of our analysis is the data listed in table XXXV of the CODATA adjustment of 2006 [4]. Since then, one result for the Planck constant has been withdrawn: the Avogadro experiment with natural silicon, $V_m$(Si)-2005 [5]. New data come from the mentioned $^{28}$Si Avogadro experiment [3], identified as Avogadro-11, the METAS watt balance experiment [6], identified as METAS-11, the new evaluation of the NPL watt balance experiments of 2006 to 2009 [7], identified as NPL-12, and the reestablished NPL watt balance experiment at NRC in Canada [8], identified as NRC-12. In total, we consider eight results from table XXXV of CODATA 2006 and the four new results. All experiments are listed in table 1, thus there are 12 measurement values and relative uncertainties for the measurand, see also figure 1. The value obtained for the Avogadro constant was converted into the corresponding $h$ value by $N_A h = 3.990\,312\,7176 \times 10^{-10}$ J s mol$^{-1}$, which has a relative standard uncertainty of $7 \times 10^{-10}$ [9].

The most commonly used estimator of a measurand in an experiment such as this is the weighted mean

$$\hat{\mu}_{\mathrm{WM}} = u^2 \sum_{i=1}^{n} \frac{x_i}{u_i^2}, \tag{1}$$

where

$$u = \left( \sum_{i=1}^{n} \frac{1}{u_i^2} \right)^{-1/2} \tag{2}$$

is its standard uncertainty. For these data, $n = 12$, $u = 1.375 \times 10^{-7} \times 10^{-34}$ J s, and $\hat{\mu}_{\mathrm{WM}} = 6.626\,069\,67 \times 10^{-34}$ J s. The relative standard uncertainty is $2.08 \times 10^{-8}$. A common check on the consistency of the data set is the chi-square test

  

**Table 1.** Measurements of the Planck constant.

| Identification | Equation in [4] | Value/$10^{-34}$ J s $x_i$ | Relative standard uncertainty $u_i/x_i$ |
|---|---|---|---|
| NPL-79 | 263 | 6.626 0729 | $1.0 \times 10^{-6}$ |
| NIST-80 | 296 | 6.626 0657 | $1.3 \times 10^{-6}$ |
| NMI-89 | 277 | 6.626 0684 | $5.4 \times 10^{-7}$ |
| NPL-90 | 282 | 6.626 0682 | $2.0 \times 10^{-7}$ |
| PTB-91 | 279 | 6.626 0670 | $6.3 \times 10^{-7}$ |
| NIM-95 | 261 | 6.626 071 | $1.6 \times 10^{-6}$ |
| NIST-98 | 284 | 6.626 068 91 | $8.7 \times 10^{-8}$ |
| NIST-07 | 288 | 6.626 068 91 | $3.6 \times 10^{-8}$ |
| METAS-11 | | 6.626 0691 | $2.9 \times 10^{-7}$ |
| NPL-12 | | 6.626 0712 | $2.0 \times 10^{-7}$ |
| NRC-12 | | 6.626 070 63 | $6.5 \times 10^{-8}$ |
| Avogadro-11 | | 6.626 070 09 | $3.0 \times 10^{-8}$ |

based on the statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{n} (\hat{\mu}_{\text{WM}} - x_i)^2/u_i^2. \tag{3}$$

Under the assumption that the data are normally distributed, independent random variables with the same mean $\mu$, and variances $u_i^2$, the $\chi^2_{\text{obs}}$ is an observed value of a chi-square random variable with $n-1$ degrees of freedom. For this data set $\chi^2_{\text{obs}} = 25.0$, and this value is very large, making the test of consistency significant at 0.05 level. This result has been generally interpreted as a proof of underestimation of the uncertainties. Of course, a large value of $\chi^2_{\text{obs}}$ can also be a consequence of one or more of the measurements being viewed as having means that are not equal to $\mu$. It is in fact impossible to reliably determine on the basis of the available data, without additional information, which of the two alternatives caused the high value of the $\chi^2_{\text{obs}}$ statistic. Reference [10] has more discussion of this interesting and important topic. The fact that the data presented in table 1 were produced by highly reputable laboratories may lead one to prefer the first explanation and endeavour to inflate the uncertainties in a sensible way. There are essentially two ways to inflate the uncertainties, one multiplicatively, described in section 2, and one additively, described in section 3. But as the second explanation is also possible we provide a method which allows for the possibility of some undetected overlooked biases in section 4.

## 2. Methods based on the Birge ratio

The method most commonly used to increase the size of the measurement uncertainties, as for example in [4], is equivalent to making the assumption that measurement $i$ is normally distributed with variance $c^2u_i^2$ and mean $\mu$. Then the expression for the consensus mean (1) remains the same, but the expression for the standard uncertainty becomes

$$u_{\text{B}} = cu.$$

The objective of the method is to estimate $c$. The usual estimate is based on the fact that under these assumptions

$$\frac{1}{c^2}\sum_{i=1}^{n}\left(\frac{x_i - \hat{\mu}_{\text{WM}}}{u_i}\right)^2 = \frac{\chi^2_{\text{obs}}}{c^2}$$

is an observed value of a chi-square random variable with $n-1$ degrees of freedom. Since the expected value of such a chi-square random variable is $n-1$,

$$E\left(\frac{\chi^2_{\text{obs}}}{n-1}\right) = \frac{(n-1)c^2}{n-1} = c^2,$$

and so $\chi^2_{\text{obs}}/(n-1)$ is an unbiased estimator of $c^2$ and $u_{\text{B}}$ can be estimated as $\tilde{u}_{\text{B}} = \sqrt{\chi^2_{\text{obs}}/(n-1)}u$. For the Planck data $\tilde{u}_{\text{B}} = 1.51u$. The final result of the analysis is therefore $\hat{\mu}_{\text{WM}} = 6.626\,069\,67 \times 10^{-34}$ J s, with relative standard uncertainty of $3.13 \times 10^{-8}$.

This method is equivalent to treating the estimate of $c$ as a constant without error. But under the same assumptions (Gaussian distribution with variance $c^2u_i^2$ and mean $\mu$) the value of $\chi^2_{\text{obs}}$ could have been different, as it is a realization of a random variable. Fortunately, it is quite straightforward to modify the Birge ratio procedure so that it accounts for this uncertainty. It requires taking the Bayesian approach [11], a line of attack in line with the *Guide to the Expression of Uncertainty in Measurement* [12], and especially its Supplement 1 [13]. (See also [14] which uses this approach to compute the gravitation constant from the 1998 CODATA data set.) The Bayesian approach requires, in addition to the assumption of a Gaussian distribution with variance $c^2u_i^2$ and mean $\mu$, a so-called prior distribution for $\mu$ and $c$. This distribution summarizes our knowledge about these two parameters before the data are taken into account. When no additional knowledge of the parameters is available, as is the case here, a so-called non-informative distribution is used. Such distributions are not uniquely determined and different criteria have been proposed for their derivation. Here we use a so-called reference prior $p(\mu, c) \propto 1/c$ (see [11] for more discussion of non-informative and reference priors, including methods of derivation). Then applying Bayes' theorem in the usual way, integrating over $c$, the posterior distribution for $\mu$ can be shown to be the three-parameter Student's $t$ distribution with $n-1$ degrees of freedom, location parameter $\hat{\mu}_{\text{WM}}$ and scale parameter $u\sqrt{\chi^2_{\text{obs}}/(n-1)}$. This density is symmetric with mean $\hat{\mu}_{\text{WM}}$ and variance $u^2\chi^2_{\text{obs}}/(n-3)$. The estimates of the consensus mean and of the standard uncertainty based on this method are $\hat{\mu}_{\text{WM}}$ and

$$u_{\text{MB}} = u\sqrt{\frac{\chi^2_{\text{obs}}}{(n-3)}},$$

and the expanded uncertainty (at a 95% level of confidence) is $u\sqrt{\chi^2_{\text{obs}}/(n-3)}t_{n-1,0.975}$. The accounting for the additional uncertainty due to the fact that the constant $c$ is not known, but only estimated, resulted in the increase of the standard uncertainty from $\sqrt{\chi^2_{\text{obs}}/(n-1)}u$ to $\sqrt{\chi^2_{\text{obs}}/(n-3)}u$.

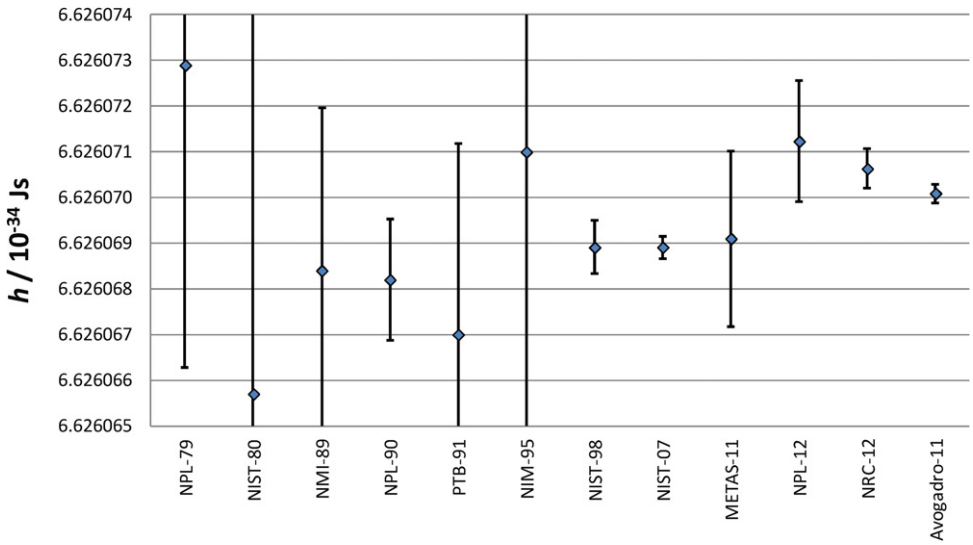


**Figure 1.** Measurements of the Planck constant $h$. The error bars denote the standard uncertainties.

Application to the Planck constant measurements yields standard uncertainty of $1.67u$, and the result $\hat{\mu}_{\mathrm{WM}} = 6.626\,069\,67 \times 10^{-34}$ J s, with relative standard uncertainty $3.46 \times 10^{-8}$. More details on this modified Birge ratio method will be reported elsewhere.

A key property of the Birge ratio adjustment methods is that the uncertainties are inflated proportionally, and therefore the usual weighted mean estimate remains unchanged. The adjustment rests on the key assumption that the underestimation of the uncertainty is proportional across the various measurements. This may or may not be true. The methods described in the next section do not make this key assumption.

## 3. Methods based on the random effects model

The random effects model [15] has a long history. It appears to have been first used by a nineteenth century astronomer Airy [16] to account for additional uncertainty in a particular estimate due to the fact that his observations were not all made under identical conditions, but instead, were recorded on several different nights with potentially different conditions. The random effects model was later rediscovered by statisticians [17, 18], and widely used ever since [10]. The model assumes that the $i$th measurement can be written as

$$x_i = \mu + \lambda_i + e_i, \tag{4}$$

where $\mu$ is the measurand, $\lambda_i$ is the so-called random effect, that is, a Gaussian random variable with mean 0 and variance $\sigma_\lambda^2$, and $e_i$, the usual measurement error term, is a Gaussian random variable with mean 0 and variance $u_i^2$. The random effect accounts for additional variability in the measurements that is not accounted for by the stated $u_i^2$. This variability is thought to be caused by possibly unknown or poorly understood factors that affect all of the measurements equally.

Under this model, a particular measurement is thought to be normally distributed with mean $\mu$ and variance $u_i^2+\sigma_\lambda^2$. Thus the individual uncertainty is increased additively by the term $\sigma_\lambda^2$. This additional variance term is estimated using the measurements. One of the natural consequences of making the uncertainty adjustment additively is that the estimate $\hat{\mu}$, given in (1), also changes.

Various procedures have been proposed in the statistical literature to estimate $\sigma_\lambda^2$. Most of them (see, for example, [19]) make use of the fact that the quantity $\sum_{i=1}^n w_i(x_i - \hat{\mu}_R)^2$ has chi-square distribution with $n-1$ degrees of freedom under model (4). Here, $w_i = 1/(u_i^2 + \sigma_\lambda^2)$, $\hat{\mu}_R = u_R^2 \sum_{i=1}^n (x_i/(u_i^2 + \sigma_\lambda^2))$ and $u_R^2 = (\sum_{i=1}^n (1/(u_i^2 + \sigma_\lambda^2)))^{-1}$. A version of this procedure which used maximum likelihood estimation was applied in [20] to compute the gravitation constant from the 1998 CODATA data set. None of the classical estimation procedures yield closed form solution for the estimator of $\sigma_\lambda^2$, and all use various simplifying assumptions or approximations. Performance of the estimators is evaluated using criteria applicable if $n$ is large, or via Monte Carlo simulation [19]. For this reason we again prefer to adopt the Bayesian paradigm as it eliminates the need for the asymptotic arguments as sample sizes of the order of 9 or 10 are not large enough.

A trade-off of course is that the Bayesian method requires a prior distribution for $\mu$ and $\sigma_\lambda^2$. As no additional information about either parameter is assumed, a non-informative reference prior $p(\mu, \sigma)$ is derived as

$$p(\mu, \sigma_\lambda) \propto \sqrt{\sum_{i=1}^{n} \frac{\sigma_\lambda^2}{(u_i^2 + \sigma_\lambda^2)^2}}.$$

This prior guarantees that a posterior distribution for $\mu$ can be obtained but its derivation requires numerical means; unfortunately no analytic expressions are available. The mean of this distribution can be used as the consensus estimate, the standard deviation as the standard uncertainty. For the Planck data, the resulting estimate is $6.626\,069\,60 \times 10^{-34}$ J s, the relative standard uncertainty is $6.68 \times 10^{-8}$. Thus the estimate is quite similar to that based on the Birge ratio but the uncertainty more than two times larger.

**Table 2.** Results of the Bayesian model averaging procedure.

| Estimate/$10^{-34}$ J s | Relative standard uncertainty | $m$ |
|---|---|---|
| 6.626 069 34 | $76.83 \times 10^{-8}$ | 1 |
| 6.626 069 46 | $22.98 \times 10^{-8}$ | 2 |
| 6.626 069 49 | $13.50 \times 10^{-8}$ | 3 |
| 6.626 069 50 | $11.09 \times 10^{-8}$ | 4 |
| 6.626 069 51 | $10.15 \times 10^{-8}$ | 5 |
| 6.626 069 55 | $9.63 \times 10^{-8}$ | 6 |
| 6.626 069 60 | $9.21 \times 10^{-8}$ | 7 |
| 6.626 069 67 | $8.69 \times 10^{-8}$ | 8 |
| 6.626 069 76 | $7.89 \times 10^{-8}$ | 9 |
| 6.626 069 87 | $6.48 \times 10^{-8}$ | 10 |
| 6.626 069 98 | $3.62 \times 10^{-8}$ | 11 |
| 6.626 069 67 | $2.08 \times 10^{-8}$ | 12 |

## 4. Methods based on the fixed effects model

It is possible that model (4) may seem overly pessimistic in the sense that one might think that at least a subset of the measurements were made with correct uncertainty evaluation. This is exactly what was done in [21]. There the assumptions of model (4) are relaxed so that a subset $\Lambda$ of size $m$ of the measurements are assumed to have the correct uncertainty assessment and so the measurements $x_i$ can be viewed as realizations of Gaussian random variables with mean $\mu$ and variance $u_i^2$. For the remaining measurements, the means are $\mu+\alpha_i$, the $\alpha_i$ are potential bias terms. These bias terms are not modelled as realizations of a common Gaussian distribution (as is done by the random effects model), rather they constitute several additional parameters to be estimated from the data. This is the so-called fixed effects model described in [15, 18].

The measurements in the set $\Lambda$ are not predetermined by the analyst; instead, all subsets of size $m$ are taken in turn and the posterior distributions of $\mu$ are obtained. The final posterior density of $\mu$ is derived via the technique called Bayesian model averaging [22] using a probability distribution which quantifies the likelihood of the various models being correct given the data. The method favours the 'most coherent' subset of size $m$; that subset will have the most influence over the value of the estimate. Table 2 gives the estimate and relative standard uncertainty for various sizes of $m$.

When $m = 12$, the procedure yields the usual weighted mean (1), and its uncertainty (2). For $m = 10$, 11 the relative standard uncertainty is smaller than for the random effects model (4). For small $m$, the uncertainty gets quite large. The results are further illustrated by figure 2.

In [21] a procedure was proposed to rule out very large values of $m$ in view of the data, but it was also explained that any method inferring the *best* value of $m$ from the data would simply be finding the most *consistent* set of data. This is not desirable as such a data set may not be the most *accurate* and a judgment of accuracy cannot be made without further information.

Application of the procedure in [21] excludes $m = 12$ (but not $m = 11$). Within the range of $m$ from 4 to 10 the uncertainty of the estimated $\mu$ decreases smoothly, and the results appear to be quite similar. (In view of the data one may therefore favour a choice of $m = 6$ or $m = 7$, say, giving similar results to the random effects model.)

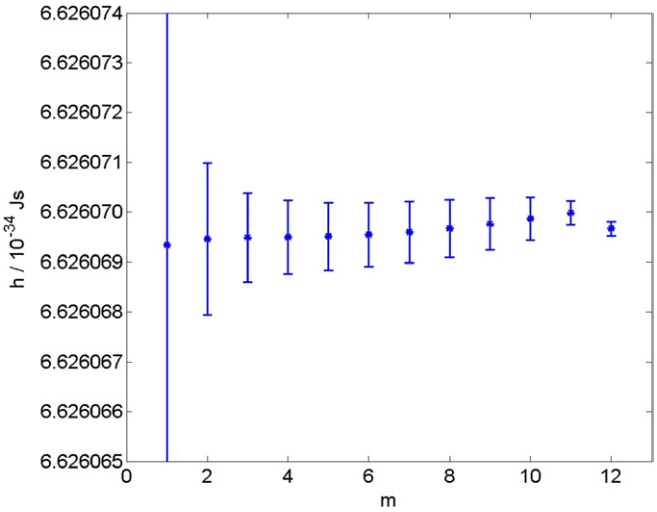


**Figure 2.** Estimates of Planck constant $h$ and standard uncertainties as a function of $m$.

The different measurements have been treated as being uncorrelated. However, sometimes some of the measurements are correlated with known correlation coefficients. Then the proposed alternative analysis methods should be extended to account for these correlations, but this is beyond the scope of this paper.

## 5. Conclusions

The estimation of a consensus value from an inconsistent set of measurements requires additional assumptions. These are best made in the form of a statistical model to either account for uncertainties that appear to be very small, or to describe overlooked effects such as possible bias terms. Which of these different sets of assumptions are the most reasonable for a particular experiment cannot be determined from the data alone, and we believe that if no additional prior information is readily available, then it is best to perform the analysis under each set and compare the results. (In this we differ from [23] which applies a set of models to measurements of the Planck constant and employs additional informative prior distributions to arrive at a preferred model.) The best situation is when several different sets of assumptions lead to similar results. This is in fact true for the Planck constant data: the random effects model analysis and the Bayesian model averaging analysis based on the fixed effects model yield very similar results (figure 3).

If in a particular case different assumptions yield different results, it is our view that either the method based on the random effects model or the fixed effects model is preferable to the adjustment based on the Birge ratio as the assumptions underlying the latter method seem the most restrictive. We see the method based on the random effects model as more restrictive than the one based on the fixed effects model [21] as there a number of the measurements are considered to be accompanied by reliable uncertainties. Only the number of

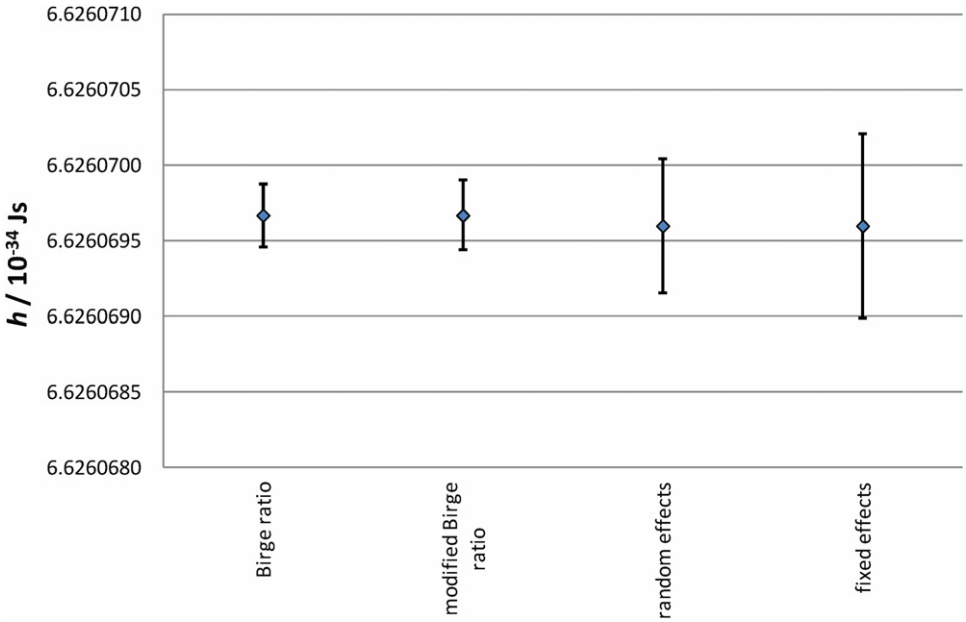


**Figure 3.** Comparison of the different methods to form a consensus estimate of $h$. For the Bayesian model averaging analysis based on the fixed effects model the result for $m = 7$ is shown. The error bars denote the standard uncertainties.

such measurements needs to be specified, not their actual identity.

## References


[1] Mills I M, Mohr P J, Quinn T J, Taylor B N and Williams E R 2006 Redefinition of the kilogram, ampere, kelvin and mole: a proposed approach to implementing CIPM Recommendation 1 (CI-2005) *Metrologia* **43** 227–46

[2] Steiner R L, Williams E R, Liu R and Newell D B 2007 Uncertainty improvements of the NIST Electronic Kilogram *IEEE Trans. Instrum. Meas.* **56** 592–6

[3] Andreas B *et al* 2011 Counting the atoms in a $^{28}$Si crystal for a new kilogram definition *Metrologia* **48** S1–13

[4] Mohr P, Taylor B and Newell D 2008 CODATA recommended values of the fundamental physical constants: 2006 *Rev. Mod. Phys.* **80** 633–730

[5] Fujii K *et al* 2005 Present state of the Avogadro constant determination from silicon crystals with natural isotopic compositions *IEEE Trans. Instrum. Meas.* **54** 854–9

[6] Eichenberger A *et al* 2011 Determination of the Planck constant with the METAS watt balance *Metrologia* **48** 133–41

[7] Robinson I A 2012 Towards the redefinition of the kilogram: a measurement of the Planck constant using the NPL Mark II watt balance *Metrologia* **49** 113–56

[8] Steele A G, Meija J, Sanchez C A, Yang L, Wood B M, Sturgeon R E, Mester Z and Inglis A D 2012 Reconciling Planck constant determinations via watt balance and enriched-silicon measurements at NRC Canada *Metrologia* **49** L8–10

[9] CODATA 2010 *CODATA Internationally Recommended Values of the Fundamental Physical Constants* http://physics.nist.gov/cuu/Constants/index.html

[10] Toman B and Possolo A 2009 Laboratory effects models for interlaboratory comparisons *Accred. Qual. Assur.* **14** 553–63

[11] Bernardo J M and Smith A F M 1994 *Bayesian Theory* (New York: Wiley)

[12] JCGM 100:2008, *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement* (Sevres, France: BIPM)

[13] JCGM 101:2008 *Evaluation of Measurement Data—Supplement 1 to the 'Guide to the Expression of Uncertainty in Measurement'—Propagation of Distributions using a Monte Carlo Method* (Sevres, France: BIPM)

[14] Dose V 2007 Bayesian estimate of the Newtonian constant of gravitation *Meas. Sci. Technol.* **18** 176–82

[15] Scheffe H 1956 Alternative models for the analysis of variance *Ann. Math. Stat.* **27** 251–71

[16] Airy G B 1861 *On the Algebraic and Numerical Theory of Errors of Observations and the Combination of Observations* (Cambridge: Macmillan)

[17] Fisher R A 1925 *Statistical Methods for Research Workers* (Edinburgh: Oliver and Boyd)

[18] Eisenhart C 1947 The assumptions underlying the analysis of variance *Biometrics* **3** 1–21

[19] Rukhin A 2009 Weighted means statistics in interlaboratory studies *Metrologia* **46** 323–31

[20] Willink R 2007 Comment on 'Bayesian estimate of the Newtonian constant of gravitation' with an alternative analysis *Meas. Sci. Technol.* **18** 2275–80

[21] Elster C and Toman B 2010 Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using Bayesian model averaging *Metrologia* **47** 113–9

[22] Hoeting J A, Madigan D, Raftery A E and Volinsky C T 1999 Bayesian model averaging: a tutorial *Stat. Sci.* **14** 382–417

[23] Mana G, Massa E and Predescu M 2012 Model selection in the average of inconsistent data: an analysis of the measured Planck-constant values *Metrologia* **49** 492–500