Compatibility verification of certified reference materials and user measurements

2014 Metrologia 51 11

(http://iopscience.iop.org/0026-1394/51/1/11)

OPEN ACCESS
IOP Publishing | Bureau International des Poids et Mesures Metrologia
Metrologia **51** (2014) 11–17 doi:10.1088/0026-1394/51/1/11

Compatibility verification of certified reference materials and user measurements

Andrew L Rukhin

Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Received 6 August 2013, revised 4 November 2013
Accepted for publication 4 November 2013
Published 6 December 2013

Abstract

A problem that frequently occurs in metrology is one of assessing compatibility of data obtained by a user laboratory with the specified values and uncertainty estimates from the certificate of analysis. The user's data are summarized by an estimated measurand value and a confidence interval, which is typically based on a repeatability standard deviation, but may include other variance or bias components. If the lab's interval and the certificate interval do not overlap, or more generally when the 'no-bias' hypothesis is rejected, the user may seek guidance on how to confirm this lack of compatibility or how to rectify it. The suggested two-stage statistical approach demonstrates a confidence interval whose width is similar to that of the certificate, and a compatibility test of guaranteed power for the given bias magnitude. Practical computationally simple formulae for each stage sample size are provided.

Keywords: compatibility testing, necessary sample size, power, Stein procedure, uncertainty, interval

1. Introduction: CRM incompatibility problem

Certified reference materials (CRMs) are well-characterized materials which are certified for one or more physical, chemical or biological properties, and are important to ensure the accuracy and compatibility of measurements. They are produced and sold in large and continually growing quantities throughout the industrialized world. 'CRMs are used for calibration, quality control and method validation purposes, as well as for the assignment of values to other materials ... and to maintain or establish traceability to conventional scales' [12]. Metrological traceability of a measurement result is often achieved by using calibrations whose quantity values are themselves traceable. However, the calibration CRM should not be subsequently used for trueness control [4].

The National Institute of Standards and Technology (NIST) produces at least 1285 individual standard reference materials (SRM is the NIST trade name for its CRMs), covering products in the major categories of chemical composition, physical properties and engineering applications and selling approximately 33 000 units per year. To certify its SRMs, NIST as well as other National Metrology Institutes reports summary statistics with associated uncertainties leading to a coverage interval. The certified value represents a resource-intensive estimate of the 'true' value of the measurand, with the interval designed to bracket this value and to indicate its uncertainty. Often users are inclined to treat the certified value as a calibration point, while the certified uncertainty is ignored. Analysts who do try to use certified uncertainty sometimes regard the interval as a target band into which the user's values must fall in order to be compliant.

The frequently cited *NIST Handbook for SRM Users* [24] does not give sufficient guidance for compatibility testing or bias removal as a corrective action. Although there is a large, rapidly expanding literature on the subject of CRM certification [14], complaints about lack of clear guidance on the use of CRMs continue, e.g. [18]. Indeed the problem of formally judging the degree of compatibility (also called 'conformance', 'trueness assessment' or 'bias determination') between a CRM certified value with associated uncertainty and a user's best estimate with its uncertainty does not seem to be fully solved.

One of the most frequently fielded queries by NIST's Measurement Services Division is related to the situation when the user's interval does not intersect the CRM interval. The usual interpretation of nonoverlapping intervals is that the measurements are not CRM compatible (i.e. the hypothesis discussed in section 3 is convincingly rejected). Non-overlap, indeed, can be taken as one of most serious indications of incompatibility, indicating the presence of possibly substantial

bias. Disjoint intervals will result in rejection of the compatibility hypothesis for virtually all existing statistical tests. More importantly, overlapping intervals do not always indicate the absence of bias.

According to the GUM [11], any significant bias should to be corrected. There is an increasing number of publications on the problem of bias removal and formal assessment of the degree of compatibility between a CRM certified value and user's best estimates, e.g. [22, 27, 28, 33]. To correct for bias, independent estimates of the bias correction uncertainty are required. Such estimates are hard or impossible to obtain in the context of just one CRM comparison where a mere replacement of the lab's mean by the certificate value should raise apprehension. This issue is touched upon in section 5.

The recent review [5] discusses more than 30 publications on the subject of measurement uncertainty and compatibility assessment over the last 15 years, among them the guidelines of ISO 10576-1 [13] and EURACHEM/CITAC [6]. See also [17]. These guides do not explicitly formulate the 'no bias' hypothesis in statistical terms. It is specified in section 3 without imposing restrictions on a lab's repeatability.

The main novel feature of ISO 10576-1 standard [13] is to recommend a two-stage stage approach to trueness assessment. If one fails to accept compatibility at stage 1, repeat the measurement and check at stage 2 by pooling measurement results from the two stages. Acceptance/rejection then depends on acceptance/rejection of the combined sample average value and its standard error. The central part of this work is the CRM experiment planning in section 3 where we follow a similar approach. Explicit formulae are given for the sample size needed in principle for each of two objectives described there: one to attain a confidence interval of the width proportional to that of CRM's interval, another to get a test which rejects compatibility with a high probability when the true bias is large.

The second stage sample sizes and the testing methodology are discussed in section 4. This procedure does not necessarily recommend to perform a compatibility test at the first stage but accomplishes one of the goals above for the combined data using the standard deviation of the first sample.

We apply the suggested methodology to two examples in section 6. However, before the methods can be explored, there are minimal requirements the user's lab must meet to show its readiness to compare their results with certified values. These issues are considered in the next section.

2. Laboratory preparation

For the purposes of applying the suggested procedure, the laboratory should have beforehand or develop an understanding of both statistical characteristics of the information contained in the certificate of an appropriate CRM and of its own measurement performance with that CRM (the estimated measurand and the uncertainty). As we will see in the next section, it is imperative that the user has a fairly good idea about the relative uncertainty of the lab's measurements with regard to that given in the CRM certificate. This can be achieved only if the measurement procedure is under statistical control.

Irrespective of calibration, traceability and quality control issues, the CRM's role is to confirm the trueness of a user's measurement results. Before using a CRM for such purpose, a lab should decide if a standard test method will be implemented. If so, the method likely has previous precision statements obtained by labs participating in its assessment. These can be compared with the CRM expanded uncertainty. However, even limited in-house validation is desirable. If there is no published method, the lab's results must be contrasted with similar off-the-shelf methods or with the work carried out for other relevant techniques. Failing that, the lab can derive some guiding characteristics from customers' specifications such as minimum allowed quantities, relative length of a specific interval, number of significant digits required in reported results, etc.

Once good repeatability is ensured, measurement precision over longer time periods and, if appropriate, among multiple analysts or different instruments should be evaluated using real samples having typical or representative analyte levels [9]. Validation studies must also establish reliable estimates of the limits of detection and quantification. Attempts to evaluate measurement trueness before fully characterizing its precision are unlikely to yield reliable statistical conclusions.

When the lab decides that it is ready to use CRMs, the next task is to choose an appropriate CRM and to employ it correctly. This topic is outside the scope of this paper. It suffices to say that the chosen CRM should have the uncertainty of certified concentrations small relative to the uncertainty for intended use. It must be reasonably matched with the customarily analysed samples and analyte concentrations. These issues are discussed, for example, in [23, 24, 32].

We focus now on statistical aspects of a CRM experiment, in particular, how to choose the number of replicates needed to detect a bias of the given magnitude, when testing the hypothesis of no bias.

3. Sample size determination: noncentral *t*-distribution

Commonly the specifications indicated in a CRM certificate provide the estimated measurand $\mu_{\rm crm}$, i.e. the certified value, and the expanded uncertainty $U_{\rm crm} = U$. Thus, $\mu_{\rm crm} \pm U_{\rm crm}$ is the uncertainty interval for the measurand μ. Traditionally used in metrology is an expansion factor 2, so that $U_{\rm crm} = 2u$ where u denotes the standard uncertainty (which commonly includes uncertainties resulting from systematic effects). For the assumed here large degrees of freedom on which the CRM interval is based, the expanded uncertainty of $(1-\alpha)100\%$ coverage interval is $z_{\alpha/2}u$, where z_β denotes the $(1-\beta)$-percentile of the standard normal distribution. When $1-\alpha = 0.95$, $z_{\alpha/2} = 1.96 \approx 2$. If the degrees of freedom are small, the factor $z_{\alpha/2}$ should be replaced by a critical value of a t-distribution which can lead to much wider intervals.

The user's replicated measurements, say, $x_1, \ldots, x_n$, are summarized by the value $\bar{x}$ which is the best estimate (typically

the sample mean) of the measurand quantity, and s which estimates the repeatability standard deviation. The number n of the measurements represents lab's sample size.

We suppose that s does not depend on $\bar{x}$ although the relationship between the sample mean and s can be quite complicated. Sometimes this independence can be approximately achieved by a suitable transformation of the x's. In particular, when s is proportional to $\bar{x}$ (as happens for some chemical measurands), the following results typically hold if x_i is replaced by the logarithmic transform, $\log x_i$.

Thus we accept here the simplest setting with x_i being a realization of a Gaussian random variable with some mean $\mu_{\rm crm} + \Delta$, where Δ represents an unknown bias, and some unknown standard deviation τ.

In this model the compatibility ('no bias') hypothesis H_0 means $\Delta = 0$. Following the tradition we denote by α the type I error or the significance level, i.e. the largest probability of false H_0 rejection. The type 2 error occurs when the null hypothesis is wrong but is not rejected. If β represents the probability of accepting H_0 when Δ is a non-zero bias, which is the type 2 error, $1 - \beta$ is called the power at Δ. A good statistical test first of all has a small significance level not exceeding α for all $\Delta = 0$, but also has a large power at least for sufficiently large $|\Delta|$ [20].

Under H_0, the ratio $\sqrt{n}(\bar{x} - \mu_{\rm crm})/s$ has a t-distribution with $\nu = n - 1$ degrees of freedom. The user confidence interval,

$$\bar{x} \pm \frac{t_{\alpha/2}(\nu)s}{\sqrt{n}},$$

is determined by $t_{\alpha/2}(\nu)$, the critical point of t-distribution with ν degrees of freedom. Thus if $\mathrm{tcdf}(t, \nu)$ is the corresponding cumulative t-distribution function, $\mathrm{tcdf}(t_{\alpha/2}(\nu), \nu) = 1 - \alpha$. The probability that this interval covers $\mu_{\rm crm} + \Delta$ is $1 - \alpha$.

Fairly often in practice this interval and the CRM interval $\mu_{\rm crm} \pm U_{\rm crm}$ do not overlap, refuting compatibility. Mathematically this fact can be expressed as

$$|\bar{x} - \mu_{\rm crm}| \geqslant U_{\rm crm} + \frac{t_{\alpha/2}(\nu)s}{\sqrt{n}}. \qquad (1)$$

When (1) holds, the lab decides to reject the compatibility hypothesis. Such a procedure is recommended by ISO 10576-1 [13]. The significance level using (1) is always smaller than α. An implication is that for small $|\Delta|$ the type 2 error is fairly large. Despite its intuitive appeal, the underpowered test (1) has a poor chance to detect a bias when it is there. It is important to realize that overlapping intervals do not imply that the lab's mean coincides with $\mu_{\rm crm}$.

Reference [30] suggests different formulations of compatibility hypothesis in metrology and provides numerical power comparisons of various procedures. In this work we concentrate on the following t-test for two reasons. First of all this test is the most commonly used technique. The second reason is that the properties of the two-stage procedure discussed in section 4 generally do not hold for other tests.

The classical t-test rejects compatibility when

$$|\bar{x} - \mu_{\rm crm}| \geqslant \frac{t_{\alpha/2}(\nu)s}{\sqrt{n}}. \qquad (2)$$

For $\Delta = 0$, the probability of false rejection using this test is exactly α. Clearly the right-hand side of (2) is always smaller than the right-hand side of (1), so each time (1) rejects, (2) rejects as well. Therefore, the probability of false acceptance under (1) is larger than that for (2), and the latter test is more powerful.

If $\Delta \neq 0$, the distribution of the ratio $\sqrt{n}(\bar{x} - \mu_{\rm crm})/s$ is known as a noncentral t-distribution with the degrees of freedom ν and the noncentrality parameter $\sqrt{n}\Delta/\tau$. In addition to controlling for the type 1 error (α), one would like to have the type 2 error (β) as small as possible. Towards this end the user may specify the minimum non-zero value for the bias, $\Delta_{\rm c}$, that is of concern. The choice of $\Delta_{\rm c}$ in practice can cause difficulties. Indeed for the bias to be deemed significant, this critical value cannot be smaller than $U_{\rm crm}$, but realistically $\Delta_{\rm c}$ should not be taken very large. We recommend to limit the values of $\Delta_{\rm c}$ to the range

$$U_{\rm crm} \leqslant \Delta_{\rm c} \leqslant 3U_{\rm crm},$$

which can be motivated by the equations below.

The larger $\Delta_{\rm c}$, the smaller is the sample size n needed to attain a given type 2 error, β, at $\Delta_{\rm c}$. The balance between α, β, n and $\Delta_{\rm c}$ can be achieved only if there is some information about the unknown τ. Indeed in our problem for a fixed α the probability of the type 2 error is a function of the noncentrality parameter $\sqrt{n}\Delta/\tau$. If τ were known, one could solve for n in the equation, type 2 error $= \beta$, to get the needed sample size n, $n \approx (z_{\alpha/2} + z_\beta)^2\tau^2/\Delta_{\rm c}^2$.

If τ is given, one can even construct a coverage interval of any width $2h$ by taking $n \approx z_{\alpha/2}^2\tau^2/h^2$. A lab may want to consider its interval having the width proportional to that of the certificate. The ratio of (expected) widths of these intervals, C_m = width(CRM interval)/width(user interval), is known as the *measurement capability index*. See [25] for a discussion of other capability characteristics.

With u representing the standard uncertainty we suggest to take $h = z_{\alpha/2}uC_m^{-1}$ for a given value of C_m. Then no matter what is τ, both of the above formulae for the sample size n coincide if

$$\Delta_{\rm c} = C_m^{-1}u\left(z_\beta + z_{\alpha/2}\right) \leqslant 4C_m^{-1}u. \qquad (3)$$

Since in the formulae for the necessary sample size τ is unknown, it must be estimated. For this purpose the comparison of τ and u is helpful. Of course one should anticipate that τ is larger than u. Take $\tau = Bu$, where the corresponding factor B, say, $1 < B \leqslant 5$, may be determined from the lab's preparatory work. To put it in a somewhat different way, let Bu be the lab's best guess about τ.

The factor B used in [30] can be described via the mentioned measurement capability index C_m, $B = [\sqrt{n}z_{\alpha/2}/t_{\alpha/2}(\nu)]C_m^{-1}$. Thus B merely is a multiple of C_m^{-1} which takes into account the error probability α and is adjusted for different sample sizes. Indeed the user's interval for very large values of $\tau/\sigma_{\rm crm}$ is practically useless. According to the rule of thumb, B should be about 2 [15]. The smallest recommended value of C_m in problems involving compliance

testing via tolerance zones is 1.3 [1, 3, 17], which requires fairly large sample sizes n. In our situation one may expect that $C_m \leqslant 1$. The lab is on equal footing with the CRM in terms of its interval width when $C_m = 1$.

Returning to the issue of controlling the type 2 error, by taking $\tau = Bu$, one gets the estimated value of the noncentrality parameter, $\sqrt{n}\Delta_c/(Bu)$, so that the numerical evaluation of the smallest n such that the inequality, type 2 error $\leqslant \beta$, becomes feasible. If Δ_c is chosen as in (3), the noncentrality parameter becomes $t_{\alpha/2}(\nu)(1 + z_\beta/z_{\alpha/2})$ with typical values between 4 and 6.

Denote by $n_m(\alpha, \beta, d)$ the minimal sample size n such that the test (2) of significance level α has the power at least $1 - \beta$ at $d = \Delta_c/\tau$. In biostatistics d is called the effect size. The modern statistical software, in particular the publicly available R-language, offers several routines to determine $n_m(\alpha, \beta, d)$ numerically for any given values of α, β and d.

Here is an R-example when sig.level $= \alpha = 0.05$, power $= 1 - \beta = 0.9$ and $d = \Delta_c/\tau = 2$:

```
library(pwr)
pwr.t.test(d=2,power=0.9, sig.level=0.05,
   type='one.sample', alternative='two.sided')
```

which produces the result

```
     One-sample t test power calculation

              n = 4.912411
              d = 2
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

According to this calculation about five observations are needed to have the type II error of test (2) equal to 0.1 when $\Delta_c/\tau = 2$.

A fairly accurate approximate formula for $n_m(\alpha, \beta, d)$ is

$$n_m(\alpha, \beta, d) \approx \left\lceil \frac{(z_{\alpha/2} + z_\beta)^2}{d^2} + \frac{z_{\alpha/2}^2}{2} \right\rceil, \tag{4}$$

where $\lceil a \rceil$ denotes the least integer that is greater or equal to a [7]. The origin of (4) is the asymptotic expansion of the noncentral t-distribution function in powers of ν^{-1} [16, chapter 31, section 5]. In the example above (4) gives the correct answer, $n_m = 5$. See also table 1.

Excel users may want to use RExcel, an add-in for MS Excel which allows to call R-functions as worksheet functions [8]. There are several web sites (http://hedwig.mgh.harvard.edu/sample_size, http://homepage.stat.uiowa.edu/~rlenth/Power, http://calculators.stat.ucla.edu/powercalc) allowing necessary sample size calculations intended mainly for clinical trials. The procedure which tests the equality of means by checking the overlap between two intervals in such studies is criticized in [31].

In the next section we will see how the user can derive an 95% confidence interval of any given length when τ is unknown. For this purpose one can employ the formula

$$n = \left\lceil BC_m \sqrt{\frac{1 + z_{\alpha/2}^2}{2}} \right\rceil, \tag{5}$$

where α is the error probability [19]. Thus, under the traditional $\alpha = 0.05$ error, $n \approx 1.55BC_m$. Equation (5) provides the optimal choice of the sample size n for $h = z_{\alpha/2}uC_m^{-1}$ with a given value of C_m as explained in [19]. This formula gives the same answer, $n = 5$, in the example above when $BC_m = 3.21$.

4. Two-stage procedure

If the CRM interval and the user interval do not overlap or, say, the compatibility hypothesis is rejected by (2), the lab may decide to follow a sequential two-stage approach to its testing recommended by [2, 10, 13]. However, none of these references specifies the necessary sample sizes. The lab's motivation may be the fact that the test (2) does not have a good power unless Δ_c/τ is fairly large [30], or it may feel that its confidence interval is very wide.

In mathematical statistics there is a method (Stein's two-stage procedure [20, p 198]) to choose the second (random) sample size m, so that when τ is unknown, the interval $\bar{X} \pm h$, $h > 0$, has a guaranteed coverage probability which is at least $1 - \alpha$, say, 95%. Here $\bar{X}$ is the total sample mean (based on both the first stage n observations and the second stage m observations). The formula for $N = n + m$ is

$$N = \max\left(n, \left\lceil \frac{s^2 t_{\alpha/2}^2(\nu)}{h^2} \right\rceil\right). \tag{6}$$

If with a given measurement capability index C_m, $h = z_{\alpha/2}uC_m^{-1}$ the lab will have its interval's width proportional to that of the CRM's interval at the expense of m additional measurements,

$$m = \max\left(\left\lceil \frac{s^2 t_{\alpha/2}^2(\nu) C_m^2}{z_{\alpha/2}^2 u^2} \right\rceil - n, 0\right). \tag{7}$$

The second sample is not needed at all when

$$s^2 \leqslant \frac{n z_{\alpha/2}^2 u^2}{C_m^2 t_{\alpha/2}^2(\nu)}.$$

Thus by claiming a small uncertainty, the lab deprives itself of the chance to re-examine its coverage interval. If the new user interval and the CRM interval still do not overlap, the lab may choose to declare its lack of compatibility or to attempt a bias correction performing further measurement runs. A motivated laboratory could perform a fully sequential sampling scheme by making measurements one at a time until for the current value of s^2, $s^2 \leqslant nz_{\alpha/2}^2 u^2/[C_m^2 t_{\alpha/2}^2(\nu)]$.

Stein's two-stage procedure also can be used in the hypothesis testing context so that for a particular bias Δ_c one can construct a test whose power for all τ is at least $1-\beta$ at this critical value. The two-stage t-test rejects the compatibility hypothesis when

$$\frac{\sqrt{N}|\bar{X} - \mu_{\text{crm}}|}{s} \geqslant t_{\alpha/2}(\nu).$$

Table 1. The necessary sample sizes when $\alpha = 0.05$, $\beta = 0.1$.

d	0.5	0.6	0.7	0.8	0.9	1	1.2	1.4	1.6	1.8	2.0	2.5	3
[23]	43	30	22	17	13	11	8	6	5	4	3	2	2
n_m	44	32	24	19	16	13	10	8	7	6	5	5	4
Equation (4)	44	32	24	19	15	13	10	8	7	6	5	4	4

If for a constant d,

$$N = N_d = \max\left(n, \left\lceil \frac{s^2 d^2}{\Delta_c^2} \right\rceil\right),$$

this procedure has the significance level α. Its smallest power, $1 - \mathrm{tcdf}(d + t_{\alpha/2}(\nu), \nu) + \mathrm{tcdf}(d - t_{\alpha/2}(\nu), \nu)$, corresponds to $\tau \to \infty$. If it is $1 - \beta$, then the test has the power at least $1 - \beta$ for all τ.

A natural choice is

$$N = N_c = \max\left[n, n_m\left(\alpha, \beta, \frac{\Delta_c}{s}\right)\right]. \tag{8}$$

with n_m from (4). Then additional observations are needed when and only when the first sample size n is smaller than $n_m(\alpha, \beta, \Delta_c/s)$, which is an estimator of the desirable theoretical quantity $n_m(\alpha, \beta, \Delta_c/\tau)$.

Another approximation via the modification of (4) suggests to take

$$N = N_a = \max\left(n, \left\lceil \frac{s^2[t_{\alpha/2}(\nu) + t_\beta(\nu)]^2}{\Delta_c^2} + \frac{t_{\alpha/2}^2(\nu)}{2} \right\rceil\right), \tag{9}$$

with the second stage sample size,

$$m = m_a = \max\left(\left\lceil \frac{s^2[t_{\alpha/2}(\nu) + t_\beta(\nu)]^2}{\Delta_c^2} + \frac{t_{\alpha/2}^2(\nu)}{2} \right\rceil - n, 0\right). \tag{10}$$

Equations (7) and (10) are approximately equal when

$$\Delta_c = C_m^{-1} u z_{\alpha/2}\left[1 + \frac{t_\beta(\nu)}{t_{\alpha/2}(\nu)}\right]\left[1 - \frac{z_{\alpha/2}^2 u^2}{2 s^2 C_m^2}\right]^{-1/2}, \tag{11}$$

which is possible only if $s^2 > 0.5 z_{\alpha/2}^2 u^2 / C_m^2$. Then about the same number of additional measurements m is required for the lab's interval to have the half-width $z_{\alpha/2} u C_m^{-1}$ as for the guaranteed power test of compatibility.

Simulations show that in terms of power the Stein procedure with N_a performs better than for N_c especially for small/medium n-values. The total sample size (9) is therefore recommended. The smallest power of this test, $1 - \mathrm{tcdf}(z_{\alpha/2} + z_\beta + t_{\alpha/2}(\nu), \nu) + \mathrm{tcdf}(z_{\alpha/2} + z_\beta - t_{\alpha/2}(\nu), \nu)$, is very close to $1 - \beta$ and considerably exceeds that of N_c in (8) for $\nu \leqslant 6$.

5. Bias uncertainty interval

The NIST Special Publication 829 [23] addresses the same issue as two previous sections, namely the design of a CRM experiment using the approximate formula

$$n_m(\alpha, \beta, d) = n_m \approx \frac{[t_{\alpha/2}(n_m - 1) + t_\beta(n_m - 1)]^2}{d^2},$$

for the necessary sample size. Here as before, β is the desired type 2 error at Δ_c, $d = \Delta_c/\tau$. This formula is obtained from the approximation of the noncentral t-random variable by the sum of a central t-random variable and the noncentrality parameter. Since n_m enters in both sides, an iterative process is required to determine its value. As the following example shows, this process may not be very accurate.

A part of table 1 in [23] for $\alpha = 0.05, \beta = 0.1$, when $d = \Delta_c/\tau$ varies from 0.5 to 3, along with exact answers obtained from the R-code and the approximate formula (4), is reproduced here. All numbers in the original table present insufficient sample sizes, while formula (4) is remarkably accurate, differing by one from the exact n_m just for two values of d ($d = 0.9$ and $d = 2.5$). The exact n_m value when $d = 2$ is 5, while the table's value 3 is far from $[t_{0.025}(2) + t_{0.1}(2)]^2/4 = 5.8$. Thus this part of [23] should not be used in practice.

Reference [23] suggests to treat u_{crm} as a fixed offset. A bias uncertainty interval for Δ is then derived from the user interval by increasing its half-width by this amount,

$$\bar{x} - \mu_{\mathrm{crm}} \pm \left[\frac{t_{\alpha/2}(\nu) s}{\sqrt{n}} + u_{\mathrm{crm}}\right].$$

The probability that the unknown bias is within these two limits is at least $1 - \alpha$, but the interval may be excessively wide for this purpose. It is closely related to the conservative test (1) as that test accepts the compatibility hypothesis if and only if the interval contains the origin.

The bias corrected interval suggested in [28] which has the end-points

$$\begin{aligned} \underline{\mu} &= \bar{x} - \max\left[\tfrac{t_{\alpha/2}(\nu)s}{\sqrt{n}} + (\bar{x} - \mu_{\mathrm{crm}}), 0\right], \\ \overline{\mu} &= \bar{x} + \max\left[\tfrac{t_{\alpha/2}(\nu)s}{\sqrt{n}} - (\bar{x} - \mu_{\mathrm{crm}}), 0\right], \end{aligned}$$

does not involve u_{crm}.

The asymmetric interval $(\underline{\mu}, \overline{\mu})$ was found to be one of the best bias removal procedures reviewed in [22]. However, this interval may not correspond to an interval defined by a coverage factor and was criticized for this reason [21].

In our problem Δ represents the short term bias which can be estimated by $\bar{x} - \mu_{\mathrm{crm}}$ with the corresponding uncertainty $\sqrt{\sigma_{\mathrm{crm}}^2 + s^2/n}$ [26]. However, for justifiable bias removal a typically missing independent estimate of this uncertainty is required. If compliance is rejected, a lab may want to pursue the bias correction by taking more measurements involving the same or different CRM and/or using other available precision information mentioned in section 2 including reproducibility conditions.

6. Two examples

We start with the following illustrative example. A laboratory purchased a CRM consisting of West Virginia coal ash, whose certified mass fraction of gallium is 58 mg kg^{-1}, with the standard measurement uncertainty 2 mg kg^{-1} evaluated on 95 degrees of freedom.

The laboratory measured this reference material as means to verify its measurement protocol, and obtained 74 mg kg^{-1} with the standard deviation 6 mg kg^{-1} evaluated on five degrees of freedom. The user 95% confidence interval which ranges from 59 mg kg^{-1} to 89 mg kg^{-1} has a small overlap with the CRM interval, 58 ± 4.

Under these circumstances in mg kg^{-1} units, $\mu_{\rm crm} = 58$, $u = 2$ while $\bar{x} = 74$, $s = 6$ and for $n = 6$, $\nu = 5$, $\alpha = 0.05$, $t_{\alpha/2}(5) = 2.57$. The t-test rejects the compatibility hypothesis as

$$\frac{\sqrt{n}|\bar{x} - \mu_{\rm crm}|}{s} = 6.53 > 2.57 = t_{\alpha/2}(5).$$

All other compatibility hypotheses tests considered in [30] also reject.

According to (7), if $C_m = 1$,

$$N = \max\left(6, \left\lceil \frac{(6 \times 2.57)^2}{4^2} \right\rceil\right) = 16.$$

Thus by taking $m = 16 - 6 = 10$ additional observations, the user would get a 95% coverage interval $\bar{X} \pm 4$. If $\bar{X} \geqslant 58 + 2.57 \times 6/\sqrt{15} \approx 61.85$, this lab may try to seek a bias correction or just state its incompatibility with this SRM.

The lab's choice for n should have been $n = 1.55B$, which suggests that in this example, $B = \tau/u > 3.8$. In this situation if $\Delta_c = 2u = 4$ is the critical bias, then according to (7) when this bias is present, additional $m = 20$ measurements for the Stein test would allow to reject H_0 correctly with probability $1 - \beta = 0.8$. The value from (11) is $\Delta_c = 6$, and the second sample size is $m = 10$, as above. If $\Delta_c = 4u = 8$, then the second sample size is reduced to $m = 5$.

We use the data for environmental SRM 1974a Organics in Mussel Tissue [29] as the second example. These data come from 16 laboratories participating in a performance-based study over a period of several years. All these labs have used $n \equiv 3$, so that $\nu \equiv 2$.

Out of 222 cases (14 compounds), the t-test rejected the compatibility hypothesis 91 times. In contrast, 41 user intervals did not overlap with the certificate interval. The three main causes, pyrene, PCB 118 and PCB 153, each contributing five instances of non-intersection, were followed by fluoranthene and 4,4′ DDT with four cases each.

As an example, consider PCB 153 with the corresponding CRM interval (145.2 ± 7.6) µg kg^{-1}. Since $t_{\alpha/2}(2) = 4.303$, lab 10 interval, (189.0 ± 10.89) µg kg^{-1}, does not overlap with the CRM interval, and the t-test rejects the compatibility hypothesis (table 2). This lab would need four additional measurements to establish a coverage interval of the same width as the CRM interval (i.e. to reduce its half-width from 10.89 to no more than 7.6.) However, this shorter interval has a poor chance to overlap with the certificate interval. The same lab would require the second sample size 10 to obtain power 0.9 when $\Delta_c = 2 \times 7.6 = 15.4$.

Table 2. The lab's intervals for PCB 153 in SRM 1974a (µg kg^{-1} units) along with the values of the additional sample size from (7) and (10) for $C_m = 1$, $\Delta_c = 15.4$, $\beta = 0.1$.

Lab	$\bar{x}$	s	(7)	(10)
10	189.00	4.38	4	10
11	184.67	5.03	6	11
12	186.50	4.95	5	11
14	182.44	2.90	0	8
16	96.47	15.26	72	45

Similarly, the lab 14 has the half-width of its reported interval, 7.20, smaller than that of CRM, so that its interval cannot be altered by a two-stage procedure. The fact that the lab 16 has a large sample standard deviation $s = 15.26$ leads to a very large additional sample size 45 for the guaranteed power test at $\Delta_c = 15.4$, and to completely unrealistic 72 new measurements for the coverage interval of half-width $2u = 7.6$. This example shows that the two-stage procedure cannot be useful in the extreme cases when the first sample standard deviation is very small or very large.

7. Conclusions

The Stein procedure offers the promise of shorter uncertainty intervals and of compatibility verification which is simultaneously more powerful and more fair to the user. This promise cannot be fulfilled by any fixed sample size statistical method. The calculations involving the necessary second sample size after (6) or (8) are sufficiently simple that end users of CRM certificates could employ them when designing a two-stage validation procedure. These formulae serve two different goals: one to attain a confidence interval of a width comparable in terms of the measurement capability index to that of the certificate, another to derive a test rejecting conformity with a high probability for the prescribed bias value. The two objectives are not incompatible; in contrast, they coincide when the critical bias value is given by (11) which can be taken as the default bias.

However, our approach needs some information about the lab's relative uncertainty with regard to CRMs. It cannot be helpful if this uncertainty is very large or very small. Neither the formula (5) nor the exact R-language calculations can be employed without certain distributional assumptions which may or may not be met in a particular situation. Effectively combining the message of the CRM certificate with the lab's data is possible only if the lab's measurement process is under statistical control. Indeed good repeatability is a precondition for any contemplated bias correction which by itself requires much more information.

Acknowledgments

The author is grateful to A Possolo for helpful suggestions, constructive critique and continual encouragement. The first example in section 6 was proposed by him. Many examples,

references and discussion were provided by the author's NIST colleagues D L Duewer and S Leigh. Useful insight by J Sieber into laboratory preparation and critical remarks of two referees are also acknowledged. Contribution of National Institute of Standards and Technology, not subject to copyright in the United States.

References

[1] ASME B89.7.3.1-2001 2002 *Guidelines for Decision Rules: Considering Measurement Uncertainty in Determining Conformance to Specifications* (New York: ASME)

[2] Christensen J M, Holst E, Olsen E and Willrich P 2002 Rules for stating when a limiting value is exceeded *Accred. Qual. Assur.* **7** 28–35

[3] Czaske M 2008 Usage of the uncertainty of measurement by accredited calibration laboratories when stating compliance *Accred. Qual. Assur.* **13** 645–51

[4] DeBiévre P, Dybkaer R, Faigelj A and Hibbert D B 2011 Metrological traceability of measurement results in chemistry: concepts and implementation (IUPAC technical report) *Pure Appl. Chem.* **83** 1873–935

[5] Desimoni E and Brunetti B 2011 Uncertainty of measurement and conformity assessment: a review *Anal. Bioanal. Chem.* **400** 1729–41

[6] Ellison S L R and Williams A (ed) 2007 EURACHEM/CITAC Guide: Use of uncertainty information in compliance assessment. www.eurachem.org/index.php/publications/guides/uncertcompliance

[7] Guenther W G 1981 Sample size formulas for normal theory *t* tests *Am. Stat.* **35** 243–4

[8] Heiberger R M and Neuwirth E 2009 *A Spreadsheet Interface for Statistics, Data Analysis, and Graphics* (New York: Springer)

[9] Hibbert D B 2007 *Quality Assurance for the Analytical Chemistry Laboratory* (Oxford: Oxford University Press)

[10] Holst E, Thyregod P and Willrich P 2001 On conformity testing and the use of two stage procedures *Int. Stat. Rev.* **69** 419–32

[11] ISO 2008 *Guide to the Expression of Uncertainty in Measurement* (Geneva, Switzerland: ISO)

[12] ISO 2000 *ISO Guide 33: Uses of Certified Reference Materials* 2nd edn (Geneva, Switzerland: International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC))

[13] ISO 2003 *ISO 10576-1: Statistical methods—Guidelines for the evaluation of conformity with specified requirements, Part 1: general principles* (Geneva, Switzerland: International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC))

[14] ISO 2006 *ISO Guide 35: Reference materials—General and statistical principles for certification* 3rd edn (Geneva, Switzerland: International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC))

[15] Instone I 1996 Simplified method for assessing uncertainties in commercial, production environment. http://metrology_forum.tm.agilent.com/easy.shtml

[16] Johnson N, Kotz S and Balakrishnan N 1995 *Univariate Continuous Distributions* vol 2, 2nd edn (New York: Wiley)

[17] Joint Committee for Guides in Metrology 2012 *JCGM 106: Evaluation of Measurement Data—The Role of Measurement Uncertainty in Conformity Assessment* www.bipm.org/utils/common/documents/jcgm/JCGM_106_2012_E.pdf

[18] Jorhem L 2004 Proper use of certified reference materials? *Accred. Qual. Assur.* **9** 507–8

[19] Lee J S and Woodroofe M 2006 A restricted minimax determination of the initial sample size in Stein's and related two-stage procedures *Random Walk, Sequential Analysis and Related Topics* (Hackensack, NJ: World Scientific Publishing) pp 44–55

[20] Lehmann E and Romano J 2003 *Testing Statistical Hypotheses* 3rd edn (New York: Springer)

[21] Lira I H and Wöger W 1998 Removing model and data non-conformity in measurement evaluation *Meas. Sci. Technol.* **9** 1010–1

[22] Magnusson B and Ellison S L R 2008 Treatment of uncorrected measurement bias in uncertainty estimation for chemical measurements *Anal. Bioanal. Chem.* **390** 201–21

[23] NIST 1992 *Use of NIST Standard Reference Materials for Decisions on Performance of Analytical Chemical Methods and Laboratories* NIST Special Publication 829, USGPO, Washington, www.nist.gov/mml/csd/inorganic/upload/NIST_SpecialPub829.pdf

[24] NIST 1993 *Standard Reference Materials Handbook for SRM Users* NIST Special Publication 260-100, USGPO, Washington www.nist.gov/srm/upload/SP260-100.pdf

[25] NIST 2010 *NIST/SEMATECH e-Handbook of Statistical Methods* www.nist.gov/itl/sed/gsg/handbook_project.cfm

[26] O'Donnell G E and Hibbert D B 2005 Treatment of bias in estimating measurement uncertainty *Analyst* **130** 721-729

[27] O'Donnell G E and Hibbert D B 2013 A study of the conditions of measurement required to evaluate bias in analytical results illustrated by the use of data from a multi-round, blind-duplicated, proficiency test *Analyst* **138** 3673–8

[28] Phillips S D, Eberhardt K R and Parry B 1997 Guidelines for expressing the uncertainty of measurement results containing uncorrected bias *J. Res. Natl Inst. Stand. Technol.* **102** 577–85 http://nvlpubs.nist.gov/jres/102/5/j25phi.pdf

[29] Poster D, Schantz M, Kucklick J, Lopez de Alda M, Porter B, Pugh R and Wise S 2004 Three new mussel tissue standard reference materials (SRMs) for the determination of organic contaminants *Anal. Bioanal. Chem.* **378** 1213–31

[30] Rukhin A L 2013 Assessing compatibility of two laboratories: formulations as a statistical testing problem *Metrologia* **50** 49–59

[31] Schenker N and Gentleman J 2001 On judging the significance of differences by examining the overlap between confidence intervals *Am. Statist.* **55** 182–6

[32] Sharpless K E and Duewer D L 2008 Standard Reference Materials for analysis of dietary supplements *J. AOAC Int.* **91** 1298–302

[33] Synek V 2005 Attempts to include uncorrected bias in the measurement uncertainty *Talanta* **65** 829–37