Metrol. Meas. Syst., Vol. 33 (2026) No. 1, pp. 1–16
DOI: 10.24425/mms.2026.158364

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS
Index 330930, ISSN 0860-8229
www.metrology.wat.edu.pl

M&MS

ON THE CONFIDENCE INTERVAL OF MEASUREMENT UNCERTAINTIES

Petr Křen

Český Metrologický Institut, Okružní 31, CZ 63800 Brno, Czech Republic (✉ *petr.kren@cmi.gov.cz*)

Abstract

Measurement uncertainty is a core term in metrology. It is widely used, but often under assumptions that are valid for a large number of measurements. However, the confidence interval of individual measurement uncertainty evaluations has not been analyzed in such depth. The confidence interval for measurement uncertainties evaluated increases as the number of measurements decreases. The paper addresses this problem, which is also important for calibration measurement capabilities as well as for evaluations of international metrological comparisons. Quantification of uncertainty is also important for new measurement and simulation methods. The paper discusses the bias and confidence interval of evaluated measurement uncertainties for normally distributed measurements and presents proposed formulae for the coverage factor for an improvement of their evaluations. The effect of correlations on measurements is also presented. The correlation estimator indicates a correlation for a small number of measurements, even though the measurements are not correlated. Therefore, a formula for the uncertainty of correlation is also presented for uncorrelated measurements. These formulae allow for an improved estimation of measurement uncertainty.

Keywords: measurement uncertainty, standard deviation, confidence interval, coverage probability, correlation.

1. Introduction

Measurement uncertainty is an integral part of the measurement result. Measurement uncertainty is often calculated using the *Guide to the Expression of Uncertainty in Measurement* (GUM) [1] and another publication [2]. Its evaluation is necessary for the decisions, measurements comparisons, or conformity assessments. Uncertainty quantification is also used in computer simulations, sensor networks, machine learning, or cognitive measurement systems. As disturbing influences are present, the mean values of measurements have a statistical character. Therefore, repeated measurements can improve the results and thus reduce the measurement uncertainty. However, the measurement uncertainty itself also has a statistical character that is often not discussed because the measurement uncertainty obtained by Type B evaluation method, which is commonly used [1], is supposed to be known *a priori*. But it is clear that if the mean as the first moment has a statistical character, then the variance and covariance as the second moments also have a statistical character as well as any other estimator such as skewness. All these moments

Article history: received February 1, 2025; revised August 8, 2025; accepted November 29, 2025; available online February 18, 2026.

together describe the position and shape of the *probability density function* (PDF), and therefore also the measurement uncertainties. The Guide [1] uses statistics for the first moment (Type A evaluation) evaluated as the experimental standard deviation of the mean. But the second moment is represented by the variance that is supposed to be known *a priori* from a PDF and obtained without a statistical analysis (Type B evaluation), whereas the covariance, which is also a second moment, is estimated statistically (Subclause 5.2). The reason for the *a priori* choice is practicality that Type B evaluation is taken from previous experience, specifications, calibration certificates, and handbooks (Subclause 4.3.1). However, since the first publication of the Guide, the ability to process large amounts of data has increased. Nevertheless, it is shown below that the variance and covariance can be estimated statistically for a smaller number of measurements with a factor that increases the estimated uncertainty of measurements, and therefore an *a priori* choice is not necessary. Moreover, the mixing of approaches with an *a priori* choice has been criticized in [3].

The evaluation of measurement uncertainty is most affected when estimates are made using a small number of measurements. An evaluation of variance can be further affected by outliers [4]. In cases with a small number of evaluations, some multiplicative factors that increase the uncertainty values are needed for a specified confidence interval. Empirical factors, named "expansion factors" in [5], for measurement uncertainties are already used by the *Committee on Data for Science and Technology* (CODATA) in the case of a small number of measurement uncertainty evaluations or due to an inconsistency in the measurements already evaluated. However, the "enlargement factor" is used for measurement uncertainties by the International Committee for Weights and Measures for the CIPM list of frequency standard values [6]. These factors do not represent the coverage factor for the measurement uncertainties in the Guide [1]. They are an extra multiplicative factor for the overall evaluation consisting of a limited number of evaluations that already include the evaluation of combined measurement uncertainties according to the Guide. In order to explore such factors, this paper presents a comparison with simulations for normally distributed measurements.

The following paper uses normal distributions of uncorrelated measurements and has the following structure. The next section deals with bias and its correction for standard deviation estimates. This section discusses these estimates on average and presents a formula for measuring uncertainty evaluation in the case of a small number of measurements. The term average is used for the arithmetic mean of estimates to avoid confusion with the arithmetic mean of measurements. Section 3 deals with the confidence interval of measurement uncertainties corresponding to their probability distribution functions. It allows us to estimate the probability of underestimation of an individual estimate of measurement uncertainty for a small number of measurements. The effect of evaluated correlations between uncorrelated measurements is discussed in Section 4. A formula for the confidence interval of the correlation is presented and compared with simulations. It is also proved that the uncertainty estimations using Student's t distribution are not a correct approach. Afterwards, the findings are summarized and discussed.

2. Average bias in standard deviation

Standard deviation is related to the second moment of a PDF. The original estimation of the standard deviation s_B is given by the equation

$$s_B = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}, \tag{1}$$

where n is the number of measurements x_i and $\overline{x}$ is the mean value. This is a downward-biased estimate of standard deviation for a small number of measurements n. Thus, it is the so-called "biased" estimate of standard deviation.

The well-known Bessel's correction uses $n-1$ instead of n in (1). It was already used in a general form by Gauss in 1823 [7] for statistical correction in the least squares method for a number of unknown parameters (to obtain the degrees of freedom) called *incognitarum* introduced in the first part of this reference, which was already written in 1821. Then, (1) changes to

$$s_U = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} \neq \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}x_i^2-\overline{x}^2}. \tag{2}$$

This estimate is sometimes called an "unbiased" estimate. However, it is also a biased estimate as will be shown below. Note that the estimator for the mean value (the first moment) must use n, and thus the bias correction for the standard deviation (obtained from the second moment) that uses $n-1$ cannot be directly used as a simple analogy of (1), but the mean value must also be corrected for this calculation.

The bias of the estimators was checked by a (pseudo)random simulation. Each measurement x_{ij} was generated using normal distribution with standard deviation s equal to one and the mean value equal to zero. We will assume for now that the mean value selected for the simulation will be equal to the obtained mean value $\overline{x}$. Then, each estimate (1) is calculated as

$$s_{Bj} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2} \tag{3}$$

and each estimate (2) is calculated as

$$s_{Uj} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}x_{ij}^2}. \tag{4}$$

Then, the average of N estimates (3) is given as

$$s_{BA}(n) = \frac{1}{N}\sum_{j=1}^{N}s_{Bj} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2} \tag{5}$$

for a given n. By analogy, the average of N estimates (4) is given as

$$s_{UA}(n) = \frac{1}{N}\sum_{j=1}^{N}s_{Uj} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}x_{ij}^2}. \tag{6}$$

The simulation result for the number of estimates $N = 10^5$ for each number of measurements n is shown in Fig. 1. A sufficient number of estimates N is chosen to avoid statistical errors in the simulations when comparing them with the analytical factors presented below. The simulation clearly demonstrates that the "unbiased" estimate (2) is upward-biased and its bias is relatively larger than the bias for the "biased" estimate (1).

The bias factor for the so-called unbiased estimates of variance and standard deviation were already presented in [8]. The same equation for this effect is present in Table E.1 in the Guide [1],

and it also uses the "unbiased" s_U. Although the estimates above do not correspond to the estimates in these references, they can be used as a hint for the bias factor formula, which is confirmed by the simulation (see Fig. 1). The bias factor f_U for s_{UA} is given as

$$f_U(n) = \frac{s_{UA}(n)}{s} = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \geq 1, \tag{7}$$

where Γ is the Gamma function. The bias factor f_B for s_{BA} is given as

$$f_B(n) = \frac{s_{BA}(n)}{s} = \sqrt{\frac{2}{n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \sqrt{\frac{n-1}{n}} f_U(n) = \frac{1}{f_U(n+1)} \leq 1. \tag{8}$$

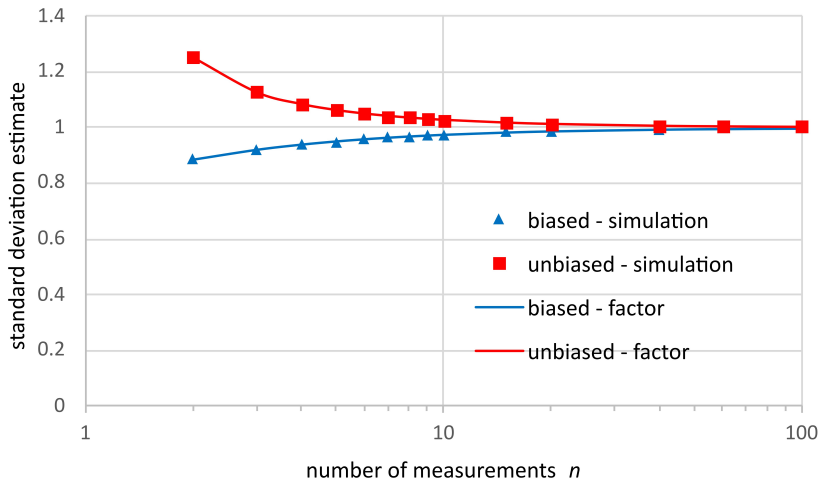

Fig. 1. Simulation results of s_{BA} and s_{UA} are shown (relatively to the unbiased s) by symbols for a given number of measurements n and the common unbiased mean value $\overline{x}$. They are compared with analytical factors f_B and f_U, respectively. The simulation and analytical estimates agree within 0.3%.

The bias factor f_B corresponds to the approximation that is numbered with the number sign #1 in [9]. The correction of bias is applied by dividing the estimate by the bias factor.

The estimates s_B and s_U use the mean value $\overline{x}$ that is common for all estimates in averages s_{BA} and s_{UA}. However, it is not a real case of estimating standard deviation. The common mean value $\overline{x}$ represents "true" statistical unbiased value that is obtained from infinite number of measurements, and thus it is inaccessible to evaluators. (The word "true" is considered redundant according to the Guide, but it helps to understand). But we can redefine (3) and (4) by using the evaluation-individual mean values $\overline{x_j}$ as

$$s_{BXj} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2} \tag{9}$$

Metrol. Meas. Syst., Vol. 33 (2026), No. 1, pp. 1–16
DOI: 10.24425/mms.2026.158364

and

$$s_{UXj} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}, \tag{10}$$

respectively. Then, the averages for simulations are

$$s_{BXA}(n) = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2} \tag{11}$$

and

$$s_{UXA}(n) = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}, \tag{12}$$

respectively. The corresponding bias factors can be found to be

$$f_{BX}(n) = \frac{s_{BXA}(n)}{s} = \sqrt{\frac{2}{n}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \approx \sqrt{\frac{n-1.5}{n}} \leq 1 \tag{13}$$

and

$$f_{UX}(n) = c_4 = \sqrt{\frac{2}{n-1}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = \frac{1}{f_U(n)} = f_B(n-1) \approx \sqrt{\frac{n-1.5}{n-1}} \leq 1, \tag{14}$$

respectively. The results are compared in Fig. 2. The bias factor (14) obtained from the "unbiased" standard deviations (10) corresponds to the factors presented in the literature [1, 7] and also to the approximation numbered #3 in [9].

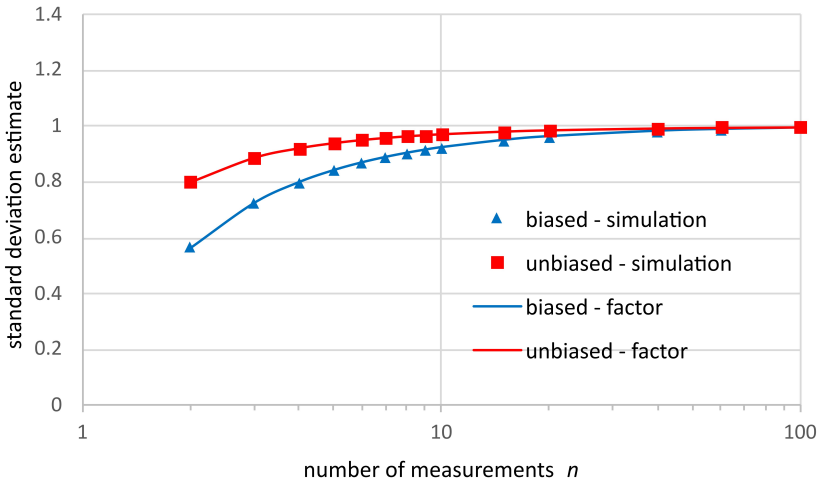

Fig. 2. Simulation results of s_{BXA} and s_{UXA} are shown (relatively to the unbiased s) by symbols for a given number of measurements n and the evaluated biased mean values. They are compared with analytical factors f_{BX} and f_{UX}, respectively. The simulation and analytical estimates agree within 0.3%.

Both the "biased" and "unbiased" standard deviations of the bias of the mean value are proportional to their coverage factor (and thus are normally distributed) and inversely proportional to the square root of n for large N. It was also confirmed by simulations that, however, are not shown here for their simplicity. The estimates available to evaluators have both mean and standard deviation biased. Thus, if the standard deviation is corrected, then there will still be a bias. A bias in the mean value $\overline{x}_j$ will produce (with the unbiased standard deviation s) lower coverage probability p_m of the "true" probability density function (without both biases) than the estimate without this bias in the mean value. The result of simulation with the bias in the mean value is shown in Fig. 3 and is compared with the following equation

$$p_m(n) = \mathrm{erf}\left(\sqrt{\frac{n}{2(n+1)}}\right), \tag{15}$$

where erf is the error function. This function was selected to describe the simulation results because the error function is the *cumulative distribution function* (CDF) of normal distribution that is shifted due to bias in the mean value. It clearly shows that a correction by factors (13) or (14) is insufficient and that the estimate is still underestimated in terms of coverage probability. Thus, the standard deviation s_M

$$s_M = \sqrt{\frac{n+1}{n}}s = \sqrt{s^2 + \left(\frac{s}{\sqrt{n}}\right)^2}, \tag{16}$$

that is corrected upward for the bias in mean value estimates should be used. The simulation confirmed (again not shown here) that the coverage probability of 68.27% with s_M does not change in average (for a large number of N) with the number of measurements n.

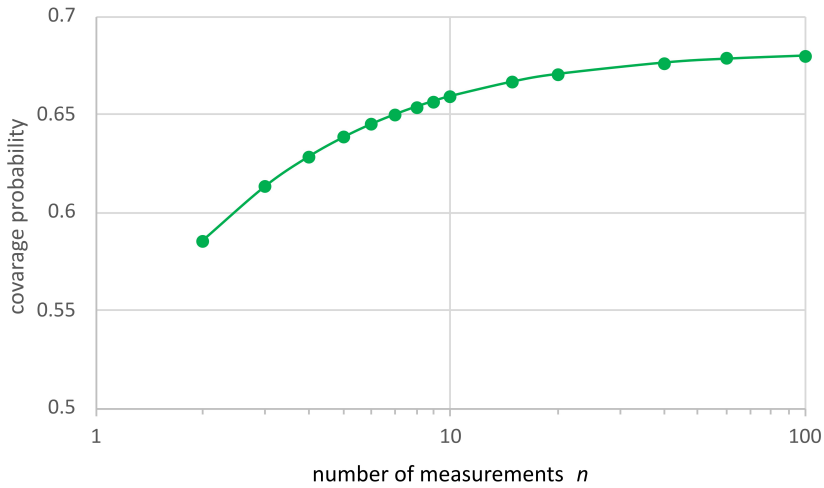

Fig. 3. Simulation results (symbols) of the average coverage probability for the mean value biased estimates with the correct standard deviation s. (16) is shown (as a line) for comparison that agrees to be relatively within 0.1%. The coverage probability of about 68.27% corresponds to the infinity number of measurements with normal distribution.

The estimate (16) for the correction of the bias in the mean value can be viewed as the combination of standard uncertainties u_C given as

$$u_C = \sqrt{u_B^2 + u_A^2}. \tag{17}$$

The standard uncertainty of Type B contribution is $u_B = s$, where the unbiased standard deviation represents normal distribution that is the result (output quantity) of a combination of input-quantity distributions. These distributions must be normal distributions, or the combination must be a result of a large number of input-quantity distributions valid for the central limit theorem. The second term in the combination (16) corresponds to the standard uncertainty of Type A, which is the experimental standard deviation of the mean [1] and is the core of the correction of standard deviation caused by the unknown bias in the mean value from a finite number of measurements. The expanded uncertainty U_C is then equal to $k_m u_C$, where k_m is the coverage factor for measurements that is valid for both the experimental standard deviations of the mean and standard deviations for normally distributed measurements because the measurements and their means are normally distributed.

In principle, the inputs and outputs of the measurement model for Type B uncertainty evaluation must have the same number of measurements n for the evaluation of possible correlations between them (using Subclause 5.2 in [1]) because a correlation can be evaluated only from pairs of measurements. These correlations can increase the measurement uncertainty, and thus measurement evaluations which suppose that uncorrelated measurements can have underestimated measurement uncertainties. Note that if individual measurements must be known for a possible covariance estimation, then they are also known for statistical estimations of variances. If the measurement model is correct, then the output uncertainties (deviations) must correspond to the combination of the input uncertainties (deviations), and thus u_A, u_B and u_C are output uncertainties and all corresponds to s. This is the case of equality $\tau = \sigma$ in [10] (where τ represents the systematic standard uncertainty and σ represents the random standard uncertainty), whereas the cases for $\tau \neq \sigma$ in [10] separate the "systematic" and "random" uncertainties, which is not a correct approach as can be seen from (16). Moreover, the Guide [1] adopts the *Recommendation INC-1* (1980) that the classification into "random" and "systematic" uncertainties should be avoided as misleading. Both the experimental standard deviations of the mean (Type A) and standard deviations (Type B) of input quantities are mathematically transferred (directly or indirectly) to the experimental standard deviations of the mean and standard deviations of the output quantities by the same measurement model with the sensitivity coefficients common for both of them.

The in-average unbiased combined standard uncertainty of input quantities for the coverage probability of 68.27% obtained from the finite number n of uncorrelated normally-distributed measurements x_i independently on the definition of standard deviation (*i.e.*, on an application of Bessel's correction) is given (using previous equations) as

$$u_C = \sqrt{\frac{n+1}{2n}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}. \tag{18}$$

The expanded combined uncertainty U_C can be obtained by multiplying (18) by the coverage factor k_m for the normal distribution of measurements. This equation can be approximated for a large number of measurements n as

$$u_C \approx \sqrt{\frac{1}{n-2.5} \sum_{i=1}^{n} (x_i - \overline{x})^2}. \tag{19}$$

The error of (19) is about 1.6% for $n = 10$, whereas (2) with Bessel's correction gives an error of –7.3% for $n = 10$ (in average), and therefore it still underestimates the evaluated uncertainty due to the insufficient factor $n - 1$. The factor $n - 2.5$ in (19) also shows that it cannot be interpreted as the degrees of freedom υ.

The application of the equations above can be illustrated using the example of a single evaluation of normally distributed data ($s = 1$) with five measurements ($n = 5$): 0.7630, –2.5351, –0.9574, 1.0314, –0.0895. Their mean value is approximately equal to –0.3575. The original estimation (1) gives a value of 1.2929, Bessel's correction (2) gives a value of 1.4455, whereas the derived one (18) gives the greatest value of 1.6846.

It is clear that the corrections presented above are not necessary for a large number of measurements n. However, the measurements must be free from different types of noise and drifts that stop the square root decrease of the Allan standard deviation with the number of measurements that is not common for longer times (see [11]), and there must be a non-correlation of the measurements that was also assumed here. Moreover, in the case of a small number of measurements n, the uncertainty evaluations presented above are valid only as the *average* of uncertainty evaluations for a large number of these uncertainty evaluations N, but not for a particular measurement uncertainty evaluation. This will also be demonstrated in the next section.

3. Distribution of standard deviation and confidence interval

The previous section dealt with the unbiased average of standard deviation estimates. However, sometimes there is a need to know from a single standard deviation estimate the interval of possible values of standard deviation with a specified probability. The probability distribution of standard deviation estimates is asymmetric [9]. Thus, the standard deviation of the standard deviation estimates is not a universally applicable estimator because it is a symmetric estimator. In light of the above, the percentile will be used instead.

The simulation used $N = 10^5$ of the "unbiased" standard deviations given by (10) and calculated their percentiles for a given number of measurements n. The simulation results are presented in Fig. 4. The simulation was also repeated (using again N estimations for $n = 2$ and $n = 100$) to estimate the simulation error. The standard deviation of the simulation results is within 1% of the values for given percentiles. The simulations show that 95% of the standard deviation

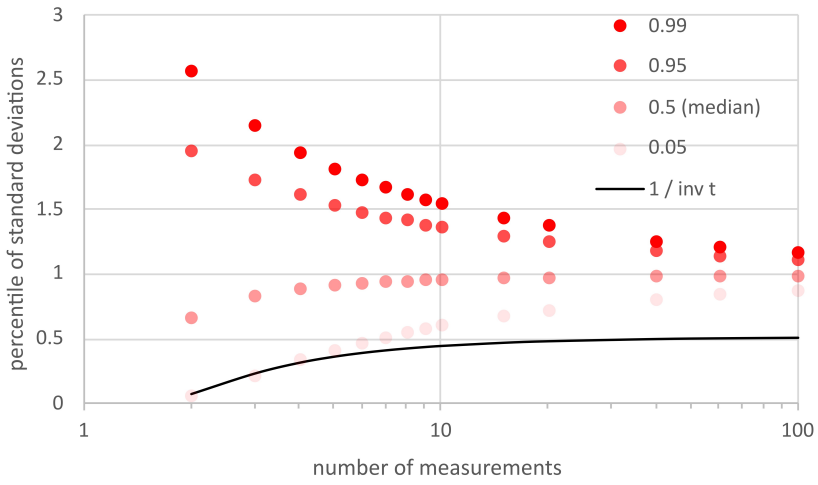

Fig. 4. Uncorrected simulation results (symbols) for percentiles of the "unbiased" standard deviations given by (10). The median is approximately (within 20%) equal to the mean from Fig. 2. The reciprocal values of the inverse Student's t distribution for probability of 0.05 are shown for comparison (line).

estimates are below approximately $2s$ for $n = 2$, which is the minimum number of measurements for which (10) is defined. But also 5% of the standard deviation estimates are below approximately $0.063s$ for $n = 2$, and thus more than 16 times underestimated. If these estimates are corrected by the bias factor for "unbiased" standard deviations (14), then the relative underestimation by more than 12 times still happens in about 5% cases for $n = 2$. It really shows that individual estimates are also much more biased than an average of large number of estimates. Then, the corrected estimate of the standard deviation for these 5% cases will be in average overestimated by about one order of magnitude. This corresponds to the fact that the confidence interval of the measurement uncertainties is relatively large for a small number of measurements n, and thus the measurement uncertainty cannot be sufficiently corrected in all cases.

A simple suggestion is to use Student's t distribution to describe the factor of underestimation discussed above. This distribution is used to estimate the coverage factor k for measurements with the degrees of freedom $\upsilon = n - 1$ [2]. Its reciprocal values are also shown for comparison in Fig. 4. For example, the evaluation for the probability 5% and $\upsilon = 1$ is approximately the reciprocal value of 12.7, which is close to factors mentioned above for $n = 2$. However, it is clear from the chart that it is not a correct formula for such estimate for all values of n.

In practice, expanded measurement uncertainties are often used (for example, the calibration measurement capabilities maintained by the Bureau International des Poids et Mesures in [12]). The most common coverage probability for measurements is equal to 95% with the coverage factor k_m of about 1.96 or about 95.46% with $k_m = 2$. The coverage probability 95% has the confidence interval with percentile endpoints 2.5% and 97.5% due to the symmetry of normal distributions. Now, we can perform a simulation (also with $N = 10^5$) that generates normally distributed data in groups of n measurements. The percentiles 2.5% and 97.5% are evaluated using linear interpolation between values. The interval between these percentiles is divided by two to obtain a single-side estimate of measurement uncertainty. The simulation results are presented in Fig. 5 and their standard deviation from repeated simulations is better than 2%. It is clear that the percentile for 95% of evaluations is also above $k_m = 1.96$ (that is valid for $n = \infty$). It means that

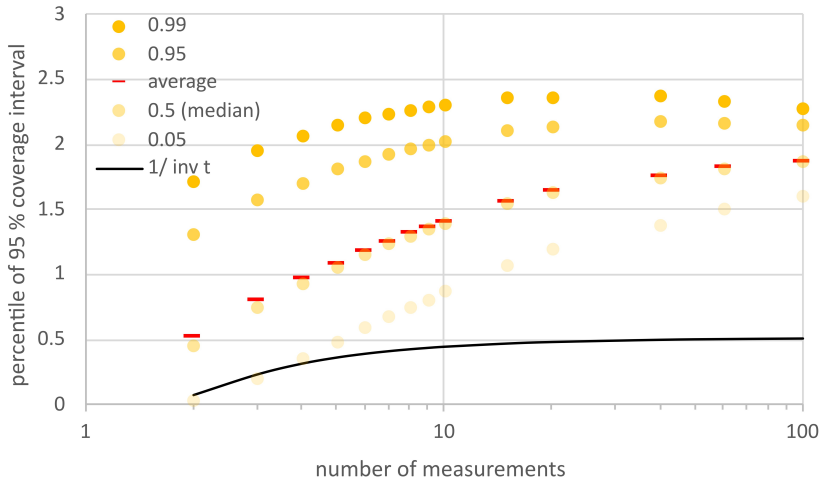

Fig. 5. Confidence interval for the coverage probability of 95% with the scale given by the unbiased standard deviation s. Symbols represent percentiles 0.05, 0.5, 0.95, 0.99 and average of these evaluated intervals for a given number of measurements n. The reciprocal values of the inverse Student's t distribution for probability 0.05 are shown for comparison (line).

the normal distribution for a finite number of measurements has tails that mimic a non-normal distribution. It could be important for various types of measurements, where the non-normal distributions are routinely observed [13]. Of course, the linear interpolation of a percentile is not an ideal approximation for small n, but it is clear from Fig. 5 that the evaluations are also relatively above $k_m = 1.96$ for $n = 100$ of measurements, where this effect of interpolation would be small. For $n \to \infty$, the fraction of estimates above k_m increases to 50%, but the estimation errors tend to zero. It has implication for the metrological comparisons, where the measurement errors relative to the estimated uncertainties are used for a validation of the calibration measurement capabilities by means of a finite number of measurements and a finite number of evaluations (participants).

Now we can compare 5% percentile for the estimates of the standard deviation and 95%-confidence interval. The uncorrected simulation results of the 5% percentile for the standard deviation presented in Fig. 4 must be corrected for the bias in the standard deviation (14). The 5% percentile of the 95%-confidence interval will be corrected by its average value in Fig. 5 to obtain the relative uncertainty of measurements uncertainties for a finite number of measurements n. These results in terms of possible underestimation of uncertainty (with respect to the average evaluation) are compared in Fig. 6. It is clear from the similarity of presented factor (agreeing within 5%) that the evaluation of percentiles by linear interpolation does not significantly affect the results of simulations. It is also clear that the factor does not directly match with the correction factor for a finite number of degrees of freedom estimated using only the inverse t^{-1} of Student's t distribution.

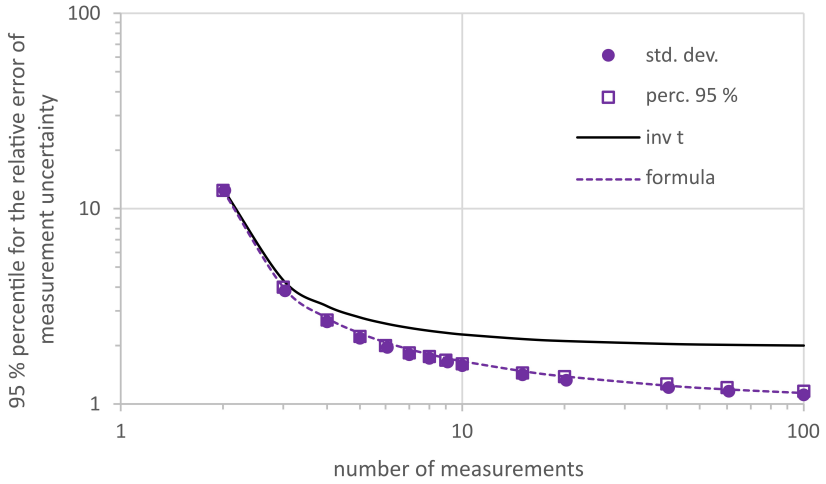

Fig. 6. The expansion factor f_u for the 5% percentile underestimation of uncertainty evaluations for the corrected standard deviations (circles) and 95%-confidence interval (squares) for a given number of measurements n. The values of the inverse Student's t distribution (solid line) and (20) (dashed line) for probability 0.05 are shown for comparison.

The simulation results obtained for the uncertainty evaluation can be approximated by the factor f_u given by the following suggested empirical formula

$$f_u \approx t^{-1}(p, n-1) - k_u + 1 + \frac{1}{4\sqrt{pn}}, \tag{20}$$

where t^{-1} is the inverse of Student's t distribution with the probability p of underestimated uncertainty evaluations ($p = 0.05$ for the example above) and the corresponding uncertainty coverage factor k_u (equal approximately to 1.96 in this example). The factor from this formula agrees with the simulations within 5% for the given example that is shown in Fig. 6.

The factor f_u that is approximated by (20) must be used to avoid (with the probability p) an uncertainty underestimation for a single given measurement uncertainty evaluation that is not an average of the large number of evaluations N described in the previous section. Thus, the combined measurement uncertainty given by (18) must be multiplied by this factor f_u (as well as multiplied by the coverage factor k_m for individual measurements) and it is valid for all input quantities of the measurement model. It is clear that this correction is not necessary for a large number of measurements n. It is claimed [2] that it is common practice to perform Type B uncertainty evaluations so that the number of measurements can be taken to be infinite. However, a low number of measurements of input quantities are performed during routine calibrations. Thus, the correction of the combined uncertainty of each input quantity by its factor f_u must be considered for a given small number of measurements and this factor can be different for a different input quantity of the measurement model.

4. Evaluated correlations of uncorrelated measurements

The simulations of uncertainties for uncorrelated measurements presented here and in [10] showed that a direct use of Student's t distribution for the corresponding measurement uncertainty evaluations of normally-distributed measurements n is an incorrect approach. Theoretical reasons follow.

Student's t distribution is the Cauchy distribution for $n = \upsilon + 1 = 2$ and it is also the normal distribution for $n = \infty$. Another example of distribution with these limiting distributions is the Voigt distribution. The Cauchy distribution has the mean value undefined for $n = 2$. However, one can calculate the mean for a pair of measurements. Thus, it is clear that the t distribution is an incorrect approach for measurement evaluations with a small number of measurements.

Standard deviation is also undefined for the Cauchy distribution, and thus a different scale must be used. The coverage factors based on the t distribution are presented in [2]. For $n = 2$, the coverage factor for the coverage probability of 95.45% is approximately 13.97 for the Cauchy distribution, whereas the coverage factor for the same probability and the normal distribution is equal to 2. However, standard deviation for normal distribution has the coverage factor equal to 1 and the corresponding coverage probability for the Cauchy distribution corresponds to the coverage factor of about 1.84. Thus, the scale of the Cauchy distribution (as well as the t distribution) is different and inconsistent with the central limit theorem that is assumed in the uncertainty evaluations for the normal distributions.

The Cauchy distribution has the property that the scale and thus the corresponding uncertainties are linearly added, whereas normal distribution has the property that the measurement uncertainties are the square root of squared uncertainties for uncorrelated measurements. However, the normally-distributed measurement uncertainties are linearly combined for fully correlated measurements ($r = 1$) [1]. Thus, there are more possibilities how to reach the linearity in uncertainty summations, including the Voight distribution.

The author of the original derivation of Student's t distribution [14] wanted to find errors in the evaluations for a small number of measurements. The author there assumes that the normally-distributed measurements are uncorrelated. However, the derivation itself assumes that an evaluated correlation (or covariance) is zero. It is a misconception that is present there because the derivation starts with standard deviation (or variance) for the error estimations in the first equation. The variance is a special case of covariance. The error in the covariance estimation is neglected in this derivation, whereas the error due to the variance should be the result of this derivation. Thus, the first equation of this derivation in [14] is already incorrect, and then the derived distribution (as it is) cannot be directly used for estimations of measurement uncertainty. The t-based methods for uncertainty estimations also cause three paradoxes: the "uncertainty paradox", the Du-Yang paradox and the Ballico paradox [15].

Generally, an estimator with a finite number of operations and a finite number of measurements that are random real numbers with infinite possibilities of values cannot result in exactly estimated zero value for this estimator. There will always be some residual value in this estimation. For example, the estimator for the mean value can be used to demonstrate this. The mean of a finite number of random real numbers is improbably equal to exactly zero. Thus, it is the reason why the evaluated correlation of a finite uncorrelated values is also not exactly zero.

The original derivation in [14] contains Bessel's correction in the second equation. It was shown above in (19) that this correction (by one degree of freedom) is not correct. Now we can see that this equation in [14] is incorrect because it reflects the first equation in [14] that is already incorrect for small n because it assumes that the evaluated correlation is zero, as it was mentioned above. The t distribution with the concept of the degrees of freedom is used for uncertainty evaluations [2]. But both, the t distribution and degrees of freedom are themselves incorrect approaches for uncertainty evaluations for the normal distributions as was already shown by simulations as well as theoretically. Of course, they can be used for other purposes, but they do not directly represent an evaluation of measurement uncertainty.

The simulation of the correlation between two sets of n uncorrelated normally-distributed measurements was carried out for a large number N of correlation evaluations. The simulation results are presented in Fig. 7. The standard deviation of the correlation coefficient u_r is given as

$$u_r = \frac{1}{\sqrt{n-1}} \tag{21}$$

within the error of simulation. The error was tested by repeating simulations with $N = 10^5$ and their results have standard deviations of less than 1%. However, the absolute value of the correlation coefficient is limited by 1, and as the distribution of the correlation coefficient is not normal distribution that extends to the infinity, the confidence interval for the normal distribution cannot be used.

The probability density function for the correlation coefficient can be found in [16]. The overview [17] presents an approximate formula obtained by the Fisher transformation that allows us to calculate the interval for a given coverage probability of the correlation coefficient. For the uncorrelated measurements, we can rewrite the limits of this interval as

$$u_{\max} = -u_{\min} \approx \tanh\left(\frac{k_r}{\sqrt{n-3}}\right), \tag{22}$$

where k_r is the z score for the normal distribution (for example, approximately equal to 1.96 for the coverage probability of 95% for the correlation coefficient). However, this equation does not allow calculations for $n < 4$. Moreover, this approximation has large errors in some cases. For example, it gives $u_{\max} \approx 0.588$ for $n = 4$ and $k_r = 0.674$, but the simulation result is $u_{\max} \approx 0.499$.

The suggested formula that allows one to calculate the interval for $n > 1$ and sufficiently agrees with simulations (with results with a standard deviation better than 1% for given percentiles) is the following

$$u_{\max} = -u_{\min} \approx \frac{1}{\sqrt{1 + \frac{n-2}{k_r^2}\left(\frac{n-2}{n}\right)^{k_r}}}. \tag{23}$$

It gives $u_{\max} \approx 0.516$ for $n = 4$ and $k_r = 0.674$, which is closer to the simulation result. Other results are shown and compared in Fig. 7. The maximal absolute value of error in the correlation value of (23) is 0.025 for all simulated data.

Metrol. Meas. Syst., Vol. 33 (2026), No. 1, pp. 1–16
DOI: 10.24425/mms.2026.158364

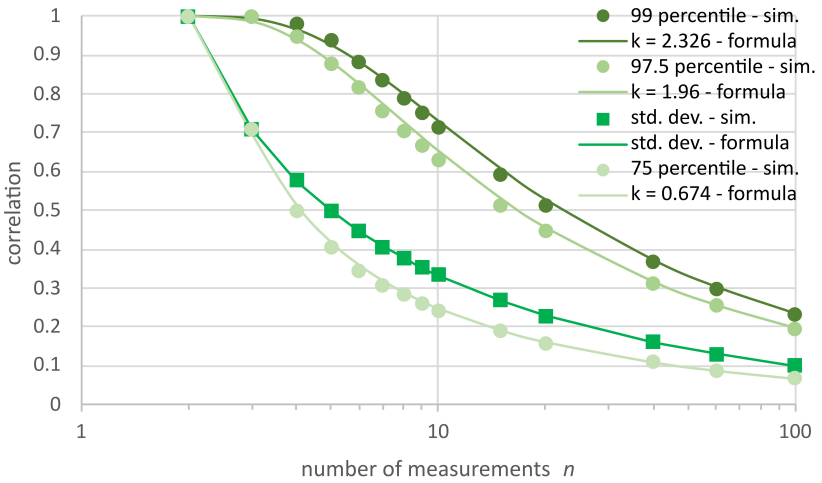

Fig. 7. Standard deviation and percentiles of simulated correlation between pairs of n uncorrelated measurements (symbols). (21) and (23) are shown for comparison (lines).

It is clear that a pair of measurements ($n = 2$) always has a correlation estimator corresponding to a fully correlated measurement, even though the generated data were produced independently as fully uncorrelated. The following is a specific example with two randomly generated sets of five measurements ($n = 5$): 0.7630, –2.5351, –0.9574, 1.0314, –0.0895 and 1.6042, –0.4567, 0.9892, –0.5920, –0.9450. The obtained non-zero correlation coefficient of 0.1745 is well below its standard deviation of 0.5 given by (21) and the endpoint of interval of 0.8815 given by (23) for $k_r = 1.96$. But, of course, some random data may have a larger correlation coefficient.

The uncertainty of the output quantity in the case of correlations r_{ij} between the input quantities with the measurement uncertainties u_i and the sensitivity coefficients c_i is given as

$$u_o^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} c_i u_i c_j u_j r_{ij} = \sum_{i=j} c_i^2 u_i^2 + 2\sum_{i>j}\sum_{j<i} c_i u_i c_j u_j r_{ij} \tag{24}$$

according to the Guide [1]. If the uncorrelated measurements are assumed, then correlations r_{ij} should not be equal to zero, but they should have the value of their uncertainty. If these contributing quantities have the same sensitivity c and measurement uncertainty u, then (24) simplifies to

$$u_o^2 = m(cu)^2 + 2(cu)^2\sqrt{\frac{m(m-1)}{2}}u_{rn} = m(cu)^2\left(1 + \sqrt{\frac{2(m-1)}{m}}u_{rn}\right), \tag{25}$$

where m is the number of input quantities and u_{rn} is the error in correlation for n measurements. The first term was added linearly because of autocorrelation, whereas the second term was added as the square root of squares because the correlations r_n are mutually random (but r_{ij} is symmetrical, and thus there is the factor of two). For $m = 2$, $n = 2$ and using (21), the uncertainty u_o is equal to $2cu$, which corresponds to the fully correlated linear combination of uncertainties for uncorrelated input quantities. Note that these input uncertainties must be already corrected for the bias (Section 2), and thus (24) is valid for the combined measurement uncertainties. They must be also corrected for the probability distribution of measurement uncertainties (Section 3) and thus evaluated as $f_u k_m u_c$ for each input quantity and then the output uncertainty is the corrected expanded uncertainty U_c.

Of course, the higher-order correlations are neglected as well as the higher central moments, which is a common approach in the measurement uncertainty evaluations. But these higher-order correlations are defined only for a sufficient number of measurements, and thus the undefined result exists if the number of measurements is lower than this order. Thus, the measurements with a low n cannot be evaluated with these higher-order terms.

5. Discussion

The results of this paper for the normal distribution can potentially help with a further update of the Guide [1] to the expression of measurement uncertainties (without *a priori* knowledge) and also with a better evaluation of metrological comparisons for which the distribution of evaluated measurement uncertainties matters. However, more work is needed to address the different probability distribution functions that are present in the measurements. It includes probability distributions such as rectangular, triangular, U-shaped distributions [18] and many others [19].

In Section 2, it was shown that the average bias in uncertainty estimations for a small number of measurements n often underestimates measurement uncertainties. The derivation of this correction formula showed that Type A uncertainty is tightly connected with Type B uncertainty evaluation obtained from the measurements, and thus the combined uncertainty with this combined correction can be used for the input quantities in the Guide. Then, the corrected combined uncertainty of input quantities is transferred to the corrected combined uncertainty of output quantities by the measurement model.

However, the probability distribution of evaluated uncertainties does not allow one to correct this bias perfectly for each uncertainty evaluation because this probability distribution function has a finite width. Moreover, the probability distribution of uncertainty evaluations is an asymmetric probability distribution for small n. Thus, a percentile level of probability cannot be easily analytically described. The simulated normal distributions have larger tails with respect to the standard deviation due to a small n and they mimic a non-normal distribution. Note that the central limit theorem is valid for a large n or m, but for a small number of measurements the probability distribution also has larger tails. It is important for 3σ or 5σ decision rules that can be applied only for a very large number of measurements. It is also expected that 5% of measurements will be discrepant with respect to the expanded measurement uncertainty in the evaluation of metrological comparisons [20], but the factor (20) suggests that they might be more discrepant. This factor is a possible further improvement of the Guide.

In the previous section, the correlations in estimates for uncorrelated normally-distributed measurements were shown. This effect is often neglected, including the Guide. The effect represents the fact that one cannot claim the non-correlation of measurements for a finite number of measurements. A fully independent measurement can be claimed (experimentally proved) only for the infinite number of measurements. The presence of a correlation in estimations of uncorrelated measurements shows that if a relation is studied, then a non-real correlation can appear (by a statistical error) for a small number of measurements. Thus, a very large number of measurements is needed, for example in frequency metrology where it is relatively easy to obtain relatively precise measurements with a given certainty.

6. Conclusions

The average bias in uncertainty estimations can be exactly corrected using derived (18) for normal distributions of uncorrelated measurements. If the “biased” estimate of the standard deviation is used, then the correction factor for bias on the scale of measurement uncertainties

is approximately 1.013 for $n = 100$. For an individual uncertainty estimation, the empirical formula (20) that corresponds to the simulations was presented in Section 3. It represents a necessary correction factor that must be applied to average evaluations of measurement uncertainties to avoid the underestimation of uncertainty of an individual estimation with a given probability. For $n = 100$ and the probability of measurement uncertainty underestimation of 5%, the correction factor for this effect is about 1.14, which is significantly more distant from 1 than the correction factor for the average bias of the measurement uncertainty evaluation equal to 1.013. The exact (21) and empirical (23) formulae were presented in order to address the issue of uncertainty in estimated correlations for uncorrelated data. Correlation coefficients of values of a few tenths are common for correlation estimations of uncorrelated data with a sample size of less than 100.

Acknowledgements

The work is funded by Institutional Subsidy for Long-Term Conceptual Development of a Research Organization granted to the Czech Metrology Institute by the Ministry of Industry and Trade.

References

[1] Joint Committee for Guides in Metrology. (2008a). Evaluation of measurement data — Guide to the Expression of Uncertainty in Measurement (JCGM 100:2008). https://doi.org/10.59161/JCGM100-2008E

[2] EA: EA-4/02 M: 2022 Evaluation of the Uncertainty of Measurement in Calibration. https://www.enac.es/documents/7020/635abf3f-262a-4b3b-952f-10336cdfae9e

[3] Willink, R. (2022). On revision of the Guide to the Expression of Uncertainty in Measurement: Proofs of fundamental errors in Bayesian approaches. *Measurement Sensors*, *24*, 100416. https://doi.org/10.1016/j.measen.2022.100416

[4] Rostron, P. D., Fearn, T., & Ramsey, M. H. (2020). Confidence intervals for robust estimates of measurement uncertainty. *Accreditation and Quality Assurance*, *25*(2), 107–119. https://doi.org/10.1007/s00769-019-01417-4

[5] Mohr, P. J., Newell, D. B., & Taylor, B. N. (2016). CODATA recommended values of the fundamental physical constants: 2014. *Reviews of Modern Physics*, *88*(3). https://doi.org/10.1103/revmodphys.88.035009

[6] Riehle, F., Gill, P., Arias, F., & Robertsson, L. (2018). The CIPM list of recommended frequency standard values: guidelines and procedures. *Metrologia*, *55*(2), 188–200. https://doi.org/10.1088/1681-7575/aaa302

[7] Gauss, C. F. (1825). *Theoria combinationis observationum erroribus minimis obnoxiae*. Henricum Dieterich. https://doi.org/10.3931/e-rara-2857

[8] Holtzman, W. H. (1950). The unbiased estimate of the population variance and standard deviation. *The American Journal of Psychology*, *63*(4), 615. https://doi.org/10.2307/1418879

[9] Hernandez, H. (2023). *Probability Distribution and Bias of the Sample Standard Deviation* (FRR 2023-02). https://doi.org/10.13140/rg.2.2.22144.51205

[10] Huang, H. (2018). A unified theory of measurement errors and uncertainties. *Measurement Science and Technology*, *29*(12), 125003. https://doi.org/10.1088/1361-6501/aae50f

[11] Riley, W.J., (2008). *Handbook of Frequency Stability Analysis*. Special Publication (NIST SP) 1065 https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=50505

[12] Bureau International des Poids et Mesures, BIPM's Key Comparison Database (KCDB). https://www.bipm.org/kcdb/

[13] Bailey, D. C. (2017). Not normal: the uncertainties of scientific measurements. *Royal Society Open Science*, *4*(1), 160600. https://doi.org/10.1098/rsos.160600

[14] Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1–25. https://doi.org/10.1093/biomet/6.1.1

[15] Huang, H. (2020). Comparison of three approaches for computing measurement uncertainties. *Measurement*, *163*, 107923. https://doi.org/10.1016/j.measurement.2020.107923

[16] Taraldsen, G. (2021). The confidence density for correlation. *Sankhya A*, *85*(1), 600–616. https://doi.org/10.1007/s13171-021-00267-y

[17] Asuero, A. G., Sayago, A., & González, A. G. (2006). The Correlation Coefficient: An Overview. *Critical Reviews in Analytical Chemistry*, *36*(1), 41–59. https://doi.org/10.1080/10408340500526766

[18] UKAS, M3003 (2024). The expression of uncertainty and confidence in measurement, Edition 6. https://www.ukas.com/wp-content/uploads/2023/05/M3003-The-expression-of-uncertainty-and-confidence-in-measurement.pdf

[19] Guthrie, W. F. (2020b). NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151) [Dataset]. In *National Institute of Standards and Technology (NIST)*. https://doi.org/10.18434/m32189

[20] Cox, M. G. (2002). Evaluation of key comparison data. *Metrologia*, *39*(6), 589–595. https://doi.org/10.1088/0026-1394/39/6/10

PetrKřen received his M.Sc. degree in physics from Charles University, Czechia, in 1998. He has been a metrologist at the Czech Metrology Institute for more than 25 years. He has more than 40 publications with nearly 500 citations. His h-index score is 11. Since 2007, he has been a delegate to the Consultative Committee for Length (CCL) – a subcommittee of the International Committee for Weights and Measures (CIPM). His research activity focuses on general metrology, optical frequency standards, interferometry, and gravimetry.