ACTA IMEKO
ISSN: 2221-870X
November 2016, Volume 5, Number 3, 16-23

Proficiency tests with uncertainty information: Analysis using the maximum likelihood method

Katsuhiro Shirono, Masanori Shiro, Hideyuki Tanaka, Kensei Ehara

National Metrology Institute of Japan, AIST Central 5, 1-1-1 Higashi, Tsukuba, Ibaraki, Japan

ABSTRACT
In this study, we report the application of the maximum likelihood method to the analysis of proficiency test data when uncertainty information is given and a reference laboratory does not exist. There are two causes that could impair the quality of analysis using the maximum likelihood method: the existence of an unknown random effect, and outliers. The conditions under which performance evaluations can be appropriately conducted are discussed in this study. To avoid serious impacts from these two causes, the maximum permissible standard uncertainty of an unknown random effect and the minimum permissible standard uncertainty of the values reported by a participating laboratory are quantified. Through simulations, the maximum and the minimum permissible standard uncertainties are found to be 0.3 and 0.5 times the intermediate magnitude of the standard uncertainty that the participants are expected to report. We believe that our proposed procedure based on these criteria is sufficiently simple to be employed in actual proficiency tests.

Section: RESEARCH PAPER

Keywords: proficiency test; E_n score; uncertainty; ISO 13528; maximum likelihood method

Citation: Katsuhiro Shirono, Masanori Shiro, Hideyuki Tanaka, Kensei Ehara, Proficiency tests with uncertainty information: Analysis using the maximum likelihood method, Acta IMEKO, vol. 5, no. 3, article 4, November 2016, identifier: IMEKO-ACTA-05 (2016)-03-04

Editor: Franco Pavese, Italy

Received March 31, 2016; **In final form** March 31, 2016; **Published** November 2016

Funding: This work was supported by a Grant-in-Aid for Scientific Research (No. 26870899) from the Japan Society for the Promotion of Science (JSPS), Japan.

Corresponding author: Katsuhiro Shirono, e-mail: k.shirono@aist.go.jp

1. INTRODUCTION

The proficiency test (PT) using an interlaboratory comparison is an effective tool to assure the quality of the measurements of calibration and testing laboratories. Participation in a PT is usually required for a laboratory to be accredited under ISO/IEC 17025:2005 [1].

For performance evaluation in a PT with information on uncertainty, a comparison with the result of a reference laboratory is implemented using the E_n score as described in ISO 13528:2015 [2], which is referred to as the E_n number in ISO 13528:2005 [3]. ISO 13528 is the standard for the statistical methods used in PTs. Given that the measurement value of Laboratory k and its expanded uncertainty are respectively x_k and U_k while those of the reference laboratory are respectively X_ref and U_ref, the E_n score for Laboratory k is defined as follows:

$$E_\mathrm{n}^{(k)} = \left(x_k - X_\mathrm{ref}\right) \Big/ \sqrt{U_k^2 + U_\mathrm{ref}^2}\ , \tag{1}$$

where superscript (k) means Laboratory k. When $|E_\mathrm{n}^{(k)}| \le 1$ and > 1, the performances of Laboratory k are evaluated as “satisfactory” and “unsatisfactory”, respectively.

Because it is difficult to designate the reference laboratory in some testing fields, performance evaluation in a PT with uncertainty information where an appropriate reference laboratory does not exist has not yet been described in ISO 13528:2015. It must be noted that when there is no reference laboratory, outliers can seriously impair the quality of the PT. On the other hand, since a PT is conducted to check the proficiency of laboratories, the possible existence of laboratories with inadequate proficiency should be taken into consideration. Thus, a robust analysis method is required.

Although no suggestion can be found in ISO 13528:2015, several analysis procedures have already been proposed for key comparison tests, which are a type of PT for national metrology institutes that are basically implemented without a specific reference laboratory. A guideline on statistical methods for key comparisons was presented by Cox [4] in 2002, in which a robust analysis referred to as Procedure B is proposed for a case where the results are inconsistent. Moreover, analysis with the largest consistent subset (LCS), also proposed by Cox [5], has been widely employed in such analysis. The LCS is the subset with the largest data size among the subsets whose consistencies are confirmed through a χ^2 test. Various other methods have also been proposed so far [6]−[10]. It is worth noting that although the statistical models employed in these proposals differ from each other, they are useful in their respective situations.

We also developed an analysis method that is robust to outliers [11], [12], which is referred to as the robust method in this paper. This method comprises two steps: (i) the detection of an unknown random effect, and (ii) performance evaluation using the local maximum likelihood (LML) method. The robust method is explained in the appendix, and a brief summary of it is given in Subsection 2.1. The advantage of this method is that it lessens the risk of performances being evaluated based on inappropriate data influenced by an unknown random effect, and allows performance evaluations to be conducted with clear statistical meaning.

In the present study, the robust method is reduced to a simpler method, which is referred to as the global maximum likelihood (GML) method. The robust method, in which LML estimators are employed, may give the impression of being difficult to implement because of its computational complexity. The approach using the GML method is explained in Subsection 2.2, and is, we believe, as simple as Algorithm A in ISO 13528:2015 Appendix C. If the conditions are clearly given, this approach can be employed in an actual PT.

Corresponding to the two steps in the robust method, conditioning is conducted by two parameters: (i) the maximum permissible standard uncertainty for a random effect, and (ii) the minimum permissible standard uncertainty for the reported values. The first parameter is quantified by examining the magnitude of a random effect that will significantly affect the quality of the PT. The second parameter is considered because, when an outlier has an extremely small uncertainty, the GML method can offer different results from the LML method. These parameters are quantified relative to the intermediate magnitude of the standard uncertainty that the participants are expected to report.

This paper is organized as follows: Section 2 provides the basic theory of the robust method and the GML method. The qualification of the parameters and the proposal of a practical procedure based on the qualified parameters are then given in Section 3. In Section 4, the procedure is applied to the data of an actual PT. A brief conclusion is presented in Section 5, and information on the robust method is provided in the Appendix.

2. ROBUST AND GLOBAL MAXIMUM LIKELIHOOD (GML) METHODS

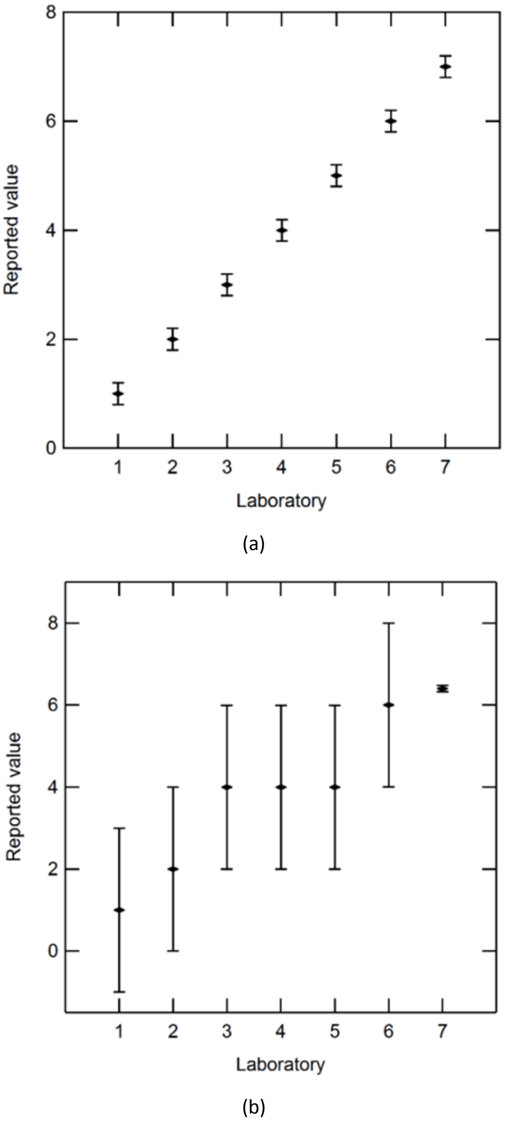
2.1. Robust method

Suppose that n laboratories participate in a PT. It is assumed that Laboratory i reports x_i and u_i as the reported value and its standard uncertainty for i = 1, 2, …, n. q_i is defined as the square of the standard uncertainty u_i.

As mentioned earlier, the two steps in the robust method are (i) the detection of an unknown random effect, and (ii) performance evaluation using the LML method. It is possible that inhomogeneity or instability of the PT items, or vagueness of the measurand, may cause a large unknown random effect and seriously impair the quality of the PT. The first step is therefore necessary. The second step is important in order to evaluate performances in a manner that is robust to outliers.

In the first step, the marginal likelihoods are computed. The marginal likelihood is, simply put, the likelihood of the model. The marginal likelihoods of the models in which an unknown random effect is not considered or is commonly considered for all n laboratories are defined as Λ_0 and Λ_n, respectively. When $\Lambda_0 < \Lambda_n$, the model with no random effect is more likely, and the data are regarded to be inappropriate for the performance evaluation. In the robust method, the marginal likelihoods of some other models with a random effect are also computed. Moreover, an estimation of the measurand is given robustly to outliers through Bayesian analysis.

Two examples are shown in Figure 1(a) and (b). In Figure 1(a), the data are largely dispersed. In this case, $\Lambda_0 < \Lambda_7$, and an unknown random effect is detected. The data are therefore inappropriate for use in the performance evaluation. On the



(a)

(b)

Figure 1. Simulated PT data: (a) (x_1, x_2, x_3, x_4, x_5, x_6, x_7) and (u_1, u_2, u_3, u_4, u_5, u_6, u_7) are given as (a) (1, 2, 3, 4, 5, 6, 7) and (0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2), and (b) (1, 2, 4, 4, 4, 6, 6.4) and (1, 1, 1, 1, 1, 1, 0.04). The error bars show the expanded uncertainties with a coverage factor of 2.

other hand, most of the data seem consistent in Figure 1(b). In this case, an unknown random effect is not detected. The estimation of the measurand is given as 4.0, which is comparable to the values of x_3 to x_5.

In the performance evaluation, the statistical model $x_i \sim \mathrm{N}(\mu, q_i + \theta_i)$ ($i = 1, 2, \ldots, n$) is considered, where μ is assumed to be the true value of the measurand in the PT, and θ_i is an additional variance. Several local maximums for the likelihood of the model can be observed. To analyse the data robustly, the combination of parameters maximizing the likelihood in which the value of μ is closest to the estimate of the measurand is focused on. Defining the LML estimators as the parameters in the combination, the symbols μ^{LML} and θ_1^{LML} to θ_n^{LML} are introduced to express the LML estimators μ and θ_1 to θ_n.

The performance evaluation for Laboratory k is then conducted to check the validity of the statistical model $x_k \sim \mathrm{N}(\mu, q_k)$ and $x_i \sim \mathrm{N}(\mu, q_i + \theta_i^{\mathrm{LML}})$ for all $i \neq k$. The specific form of the statistic is given in the appendix as the extended E_n score. When $|E_\mathrm{n}^{(k)}| \leq 1$ and > 1, the performances of Laboratory k are evaluated respectively as "satisfactory" and "unsatisfactory" as described in connection with (1).

The extended E_n scores obtained using the robust method for the data in Figure 1(b) are given in Table 1. The extended E_n scores of x_3 to x_5, whose values are close to the estimation of the measurand, are 0.0, and their performances are evaluated as "satisfactory". On the other hand, the extended E_n score of the apparent outlier of x_7 is 2.4, and the performance is evaluated as "unsatisfactory". Thus, it is found that the robust method is not significantly influenced by outliers.

2.2. Global maximum likelihood (GML) method

Although the robust method is flexibly applicable, the computation may be complicated. Therefore, it is meaningful to provide a simpler method. Analysis using the GML estimator is proposed here. In this section, n, x_i, u_i, and q_i are defined as in Subsection 2.1.

The statistical model $x_i \sim \mathrm{N}(\mu, q_i + \theta_i)$ ($i = 1, 2, \ldots, n$) is considered in this method as in the robust method. This model is actually identical to Model III proposed by Willink [8]. In Willink's paper, the GML estimator of θ_i, θ_i^{GML}, is given as $\theta_i^{\mathrm{GML}} = \max\{|x_i - \mu^{\mathrm{GML}}|^2 - q_i, 0\}$, where μ^{GML} is the GML estimator of μ. Moreover, Willink showed that μ^{GML} can be given through the minimization of the following quantity $Q(\mu)$:

$$Q(\mu) = \sum_{i=1}^{n} \left\{ \log(\phi_i) + \frac{(\mu - x_i)^2}{\phi_i} \right\}, \tag{2}$$

where $\phi_i = \max\{|x_i - \mu|^2, q_i\}$.

We apply the same concept as in the robust method to the performance evaluation. From an analogy to $E_\mathrm{n}^{(k)}$ in the robust method given in (A5) in the appendix, the following extended E_n score is computed for the performance evaluation of Laboratory k with $\phi_i^{\mathrm{GML}} = \theta_i^{\mathrm{GML}} + q_i$:

$$E_n^{(k)} = \frac{x_k - \left(\sum_{i \neq k}^{n} 1/\phi_i^{\mathrm{GML}}\right)^{-1} \left(\sum_{i \neq k}^{n} x_i/\phi_i^{\mathrm{GML}}\right)}{2\sqrt{u_k^2 + \left(\sum_{i \neq k}^{n} 1/\phi_i^{\mathrm{GML}}\right)^{-1}}}, \tag{3}$$

where $\Sigma_{i \neq k}^{n}$ denotes the summation for $i = 1, 2, \ldots, k - 1, k + 1, \ldots, n$. This statistic is analogic to the concept of the "exclusive statistics" [13], [14]. When $|E_\mathrm{n}^{(k)}| \leq 1$ and > 1, the performances of Laboratory k are evaluated as "satisfactory" and "unsatisfactory", respectively. This analysis method is referred to as the GML method in this study.

Since no check of an unknown random effect is conducted in this method, the GML method is not appropriate when an unknown random effect has a serious impact on the PT data. Therefore, the GML method should not be applied to the data shown in Figure 1(a) without any quantitative consideration of the unknown random effect.

Moreover, μ^{GML} and μ^{LML} may differ from each other when an extremely small uncertainty is associated with an outlier. When the GML method is applied to the data in Figure 1(b), the extended E_n scores are reported as shown in Table 1. The computed extended E_n score of Laboratory 7 is 0.8, and the performance is evaluated as "satisfactory". This unnatural result occurs due to the extremely small uncertainty of x_7.

However, we believe that the GML method is sufficiently simple to be employed in an actual PT analysis if certain conditions are satisfied. Specifically, the method is applicable when it is confirmed that the uncertainty of an unknown random effect is negligibly small, and no extremely small standard uncertainties are reported by the participating laboratories. In Section 3, the conditions are discussed in quantitative terms and a practical procedure is proposed.

Table 1. Extended E_n scores computed using the robust method and the GML method.

Laboratory (k)	$E_\mathrm{n}^{(k)}$ in the robust method	$E_\mathrm{n}^{(k)}$ in the GML method
1	− 1.4	− 2.7
2	− 1.0	− 2.2
3	0.0	− 1.2
4	0.0	− 1.2
5	0.0	− 1.2
6	0.9	− 0.2
7	2.3	0.8

3. CONDITIONING FOR APPLICATION OF THE GLOBAL MAXIMUM LIKELIHOOD APPROACH

3.1. Concept of the maximum and the minimum permissible uncertainties

In this section, we focus on the two factors that might cause a problem in the performance evaluation proposed in Subsection 2.2: (i) an unknown random effect, and (ii) outliers. The first factor can be neglected when the random effect is much smaller than the uncertainty of the reported values, while the second factor can be neglected when no extremely small uncertainty is reported.

Once it is clarified how large an uncertainty will be reported by the participants in a PT, criteria to avoid the effects of these factors can be examined. The standard uncertainty that is expected to be averagely reported by the participants in a PT is referred to as the expected standard uncertainty in this study, and expressed as u_{exp}. It is assumed that the expected standard uncertainty can be roughly determined based on technical knowledge in advance of the PT.

The maximum permissible standard uncertainty of an unknown random effect and the minimum permissible standard uncertainty of the reported values are introduced to avoid the influences of the random effect and outliers, respectively. These are discussed in Subsections 3.2 and 3.3 based on the assumption that the expected standard uncertainty u_{exp} is appropriately given. In Subsection 3.4, a practical procedure is proposed based on these discussions.

3.2. Influence of the random effect

Only inhomogeneity and instability are considered as sources of an unknown random effect in this study. These random effect sources tend to have a serious impact on the analysis of a PT. It is assumed that it has been confirmed that the definition of the measurand, including the choice of the measurement methods, will not have a serious influence.

Defining σ_{rnd} as the standard deviation of the random effect, the case in which σ_{rnd} is estimated to be 0.3 × u_{exp} is considered. The criterion that the standard deviation of the inhomogeneity must not exceed 0.3 × σ_{P}, where σ_{P} is the standard deviation for the proficiency assessment, is given in ISO 13528:2015. Regarding the instability, almost the same criterion is used. This criterion is applied by extension to a case with uncertainty information in this study. Since σ_{P} could be interpreted as the average value of the dispersion of the reported values, it seems appropriate to replace σ_{P} with u_{exp}.

To characterize these criteria quantitatively, letting $\Phi^{-1}(.)$ be the inverse of the cumulative standard normal probability function, it is considered that x_i is derived from $\mathrm{N}(\mu, q_i + \sigma_{\mathrm{rnd}}^2)$ and the data are given as follows:

$$x_i = \sqrt{\frac{1+0.3^2}{V}} \times \Phi^{-1}\left(i/(n+1)\right), \tag{4}$$

$q_i = u_{\mathrm{exp}} = 1$ for $i = 1, 2, \ldots, n$,

where $z_i = \Phi^{-1}(i/(n+1))$, and $V = \Sigma_{i=1}^{n} (z_i - \Sigma_{i=1}^{n} z_i/n)^2/(n-1)$. Using these data, σ_{rnd}^2 is unbiasedly estimated to be 0.3^2. It should be noted that this sequence of x_i does not necessarily mean that x_i is truly derived from $\mathrm{N}(0, 1 + 0.3^2)$, and other statistical models including $x_i \sim \mathrm{N}(\mu, q_i + \theta_i)$ ($i = 1, 2, \ldots, n$) employed in the robust method might yield the same sequence of x_i. In Figure 2, the dispersion of data when $n = 30$ is shown.

Consequently, the unknown random effect is undetected when the data are given by (4). The marginal likelihoods of the models in which an unknown random effect is considered for no and all laboratories, Λ_0 and Λ_n, are computed for the cases with n = 2, 5, 10, 20, 100, and 200. Figure 3(a) shows Λ_0 and Λ_n, implying that Λ_0 is always larger than Λ_n. Thus, it is concluded based on the discussion in Subsection 2.1 that no unknown random effect is detected.

Since the result could change if different data were provided, the property is checked with more dispersive data. The following data are considered:

$$x_i = \sqrt{\frac{1+0.5^2}{V}} \times \Phi^{-1}\left(i/(n+1)\right), \tag{5}$$

$q_i = u_{\mathrm{exp}} = 1$ for $i = 1, 2, \ldots, n$.

Figure 3(b) shows Λ_0 and Λ_n, and $\Lambda_0 > \Lambda_n$ for all n as well. This means that the unknown random effect is undetectably small even when σ_{rnd} is roughly estimated to be 0.5 × u_{exp}.

Therefore, it is concluded that 0.3 × u_{exp} could be a strong candidate for the maximum permissible standard uncertainty. In this discussion, σ_{rnd} is unbiasedly estimated using the PT data. In ISO 13528:2015, on the other hand, inhomogeneity and instability are basically evaluated independently from the PT data. However, since the random effect is undetectably small even with more dispersive data, it can be said to be conservative to set the maximum permissible standard uncertainty as 0.3 × u_{exp}, irrespective of the method used for the estimation of σ_{rnd}.

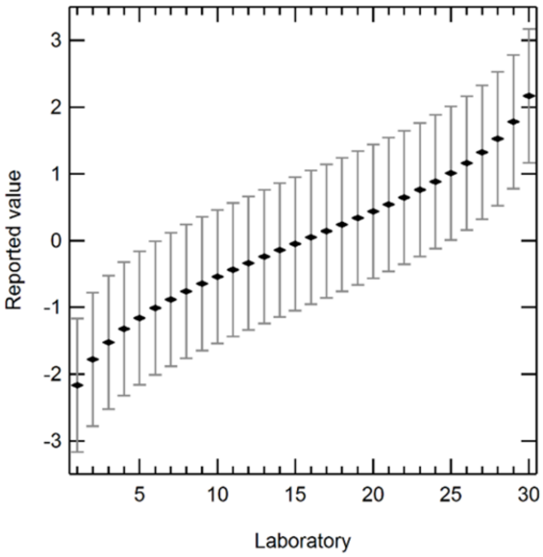


Figure 2. Simulated PT data given by (4) when n = 30. The error bars show the expanded uncertainties with a coverage factor of 2.

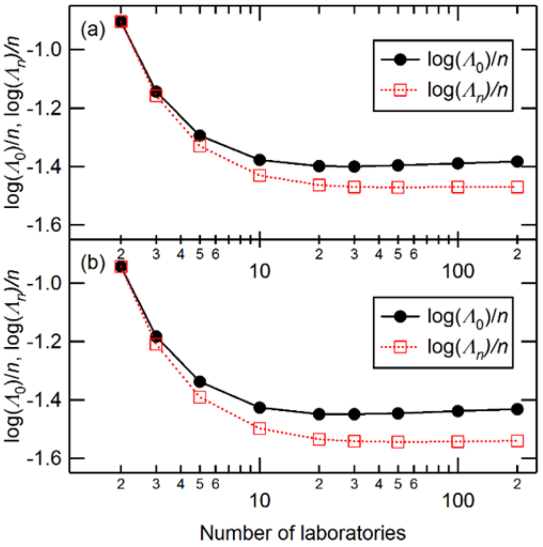


Figure 3. Computed logarithmic marginal likelihoods per number of laboratories for the data when n = 2 to 200 using the data given in the equations of (a) (4) and (b) (5).

3.3. Sensitivity to outliers

We recommend that the GML method be employed only when the number of the laboratories with $|E_{\mathrm{n}}| \leq 1$ is 10 or more. When the number of participants is small, the discrimination of outliers from the other values is technically difficult. Such discrimination is, however, necessary in the case of the GML method. In the following discussion, the GML method is characterized through simulated data with an outlier and 10 other data. Cases with a smaller data size are not taken into consideration.

We believe that 0.5 × u_{exp} is appropriate as the minimum permissible standard uncertainty mentioned in subsection 3.1. It is considered that the uncertainty should be larger than 1/√10 × $u_{\mathrm{exp}} \approx 0.32 \times u_{\mathrm{exp}}$, because the weighted mean of the 10 data with the standard uncertainty of u_{exp} is given as 1/√10 × u_{exp}. When $u_i < 1/\sqrt{10} \times u_{\mathrm{exp}}$, the value x_i can have a strong impact on the determination of the GML estimator. Thus, the minimum uncertainty of 0.5 × u_{exp} seems a possible choice. The robustness of the GML method with this value is consequently examined in this subsection.

In discussing the property of the GML method in quantitative terms, it is assumed that $u_{\mathrm{exp}} = 1$, and x_1 to x_{10} are considered to be given as follows:

$$x_i = \sqrt{1/V} \times \Phi^{-1}(i/11) \; , \qquad (6)$$

$q_i = u_{\mathrm{exp}}{}^2 = 1$, where V is defined in Subsection 3.2. In addition to these data, x_{11} and q_{11} are basically determined differently so as to be characterized as outliers. See Figure 4 as an example with $x_{11} = 2.0$ and $q_{11} = 0.5^2$. Examples with $q_{11} = 0.3^2$ and 0.5^2 are described in this study; the analysis with $q_{11} = 0.3^2$ is conducted for comparison. If the simulation with the smaller minimum permissible standard uncertainty of 0.3 × u_{exp} provides a result that is robust to the outliers, it can be said that setting the minimum permissible standard uncertainty as 0.5 × u_{exp} is an adequately conservative choice.

Consequently, the criterion of a minimum permissible standard uncertainty of 0.5 × u_{exp} seems appropriate. When $\mu^{\mathrm{LML}} = \mu^{\mathrm{GML}}$, the identical extended E_{n} scores are given in both the LML and the GML methods for all of the laboratories. Figure 5(a) and (b) show a comparison between μ^{LML} and μ^{GML} for $q_{11} = 0.3^2$ and 0.5^2, respectively. μ^{LML} and μ^{GML} are in perfect agreement with each other under the conditions of both $q_{11} = 0.3^2$ and $q_{11} = 0.5^2$. Even in the case of $q_{11} = 0.3$, the results are not significantly contaminated by the existence of outliers. Thus, the validity of the criterion of 0.5 × u_{exp} is confirmed.

It should be noted that the participating laboratories must agree in advance to a PT in which standard uncertainty less than the minimum permissible standard uncertainty is not usually reported, and will not be reported in the PT. This is because the measurements for a PT must be conducted under the usual experimental conditions. It is not recommended that a laboratory that usually reports an uncertainty of less than 0.5 × u_{exp} reports an uncertainty of 0.5 × u_{exp} or more only for the purpose of this PT.

3.4. Proposed procedure using the GML method

Based on the discussions in Subsections 3.2 and 3.3, we suggest the following four conditions for application of the GML method: (i) no unknown random effect other than inhomogeneity and/or instability of the PT items exists, (ii) the random effect caused by the inhomogeneity and/or instability has a standard deviation of less than 0.3 × u_{exp}, (iii) the minimum permissible standard uncertainty from the laboratories is 0.5 × u_{exp}, and (iv) 10 or more laboratories report data for which the performance is evaluated as "satisfactory".

The following procedure entailing 11 steps is proposed to realize the above conditions:

1. It is confirmed that there is no or practically negligible unknown uncertainty in the definition of the measurand, including the choice of the measurement method.
2. The PT provider determines the expected uncertainty, u_{exp}, from the existing technical knowledge.

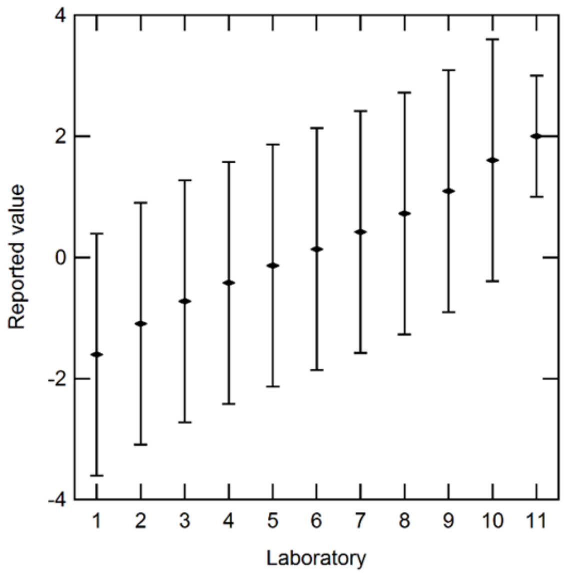


Figure 4. Simulated PT data given by (7) when $x_{11} = 2.0$ and $q_{11} = 0.5^2$. The error bars show the expanded uncertainties with a coverage factor of 2.

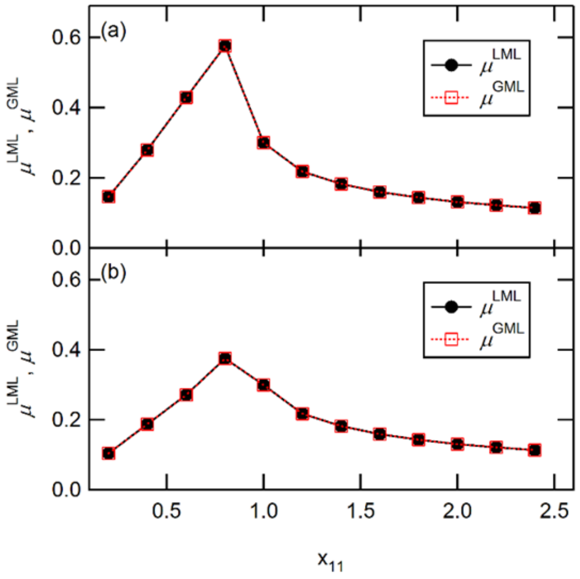


Figure 5. Computed GML and LML estimators with $q_{11} = 0.3^2$ and 0.5^2 as a function of x_{11}. x_1 to x_{10} yielded by (7).

3. The PT provider checks that the standard deviation of the random effect caused by the inhomogeneity and/or instability of the PT items is less than 0.3 × u_{exp}.
4. All of the participants agree that a standard uncertainty of less than 0.5 × u_{exp} is not usually reported, and will also not be reported in the PT.
5. The PT is implemented and the data of x_i and u_i (i = 1, 2, …, n) are obtained.
6. It is confirmed that a representative value (e.g., the median) of the reported standard uncertainties is adequately close to the expected uncertainty.
7. $$\mu^{\mathrm{old}} = \underset{x_i \in x_1, \ldots, x_n}{\arg\min} \left[\sum_{i=1}^{n} \left\{ \log(\phi_i) + \frac{(\mu - x_i)^2}{\phi_i} \right\} \right].$$

8. $$\mu^{\mathrm{GML}} = \frac{\sum_{i=1}^{n}\left\{x_i / \max\left(q_i, \left(x_i - \mu^{\mathrm{old}}\right)^2\right)\right\}}{\sum_{i=1}^{n}\left\{1/ \max\left(q_i, \left(x_i - \mu^{\mathrm{old}}\right)^2\right)\right\}}.$$

9. If $|\mu^{\mathrm{GML}} - \mu^{\mathrm{old}}| > \varepsilon[\Sigma_{i=1}^{n} \{1/\max(q_i, (x_i - \mu^{\mathrm{GML}})^2)\}]^{-1/2}$, where ε is a small value like 10^{-3}, $\mu^{\mathrm{old}} = \mu^{\mathrm{GML}}$ and go to Step 8.

10. For $k = 1, 2, \ldots, n$, $E_{\mathrm{n}}^{(k)}$ is computed through

$$E_n^{(k)} = \frac{x_k - \left(\sum_{i\neq k}^{n} 1/\phi_i^{\mathrm{GML}}\right)^{-1}\left(\sum_{i\neq k}^{n} x_i/\phi_i^{\mathrm{GML}}\right)}{2\sqrt{u_k^2 + \left(\sum_{i\neq k}^{n} 1/\phi_i^{\mathrm{GML}}\right)^{-1}}},$$

where $\phi_i^{\mathrm{GML}} = \max(q_i, (x_i - \mu^{\mathrm{GML}})^2)$.

11. It is confirmed that 10 or more laboratories have a performance of $|E_{\mathrm{n}}^{(k)}| \leq 1$.

For Steps 1, 3, 4, 6, and 10, if the conditions are not satisfied, the robust method should, in principle, be implemented instead of the GML method. However, in Step 4, if a laboratory requests permission to report a smaller standard uncertainty than the minimum permissible uncertainty, such a request may be accepted only when the laboratory has sufficient technical evidence.

In Step 7, the initial value for the estimation of μ^{GML} is determined as x_i with which $Q(x_i)$ in (2) is the smallest among all $Q(x_i)$. Although it is not mathematically assured that the GML estimator can be obtained through this estimation, we have not found a case in which the GML estimator is not given by this determination of the initial value.

The chief advantage of the GML method is that the algorithm is sufficiently simple to be employed in an actual PT; or in other words, to be incorporated into ISO 13528:2015. In ISO 13528:2015, several algorithms for a PT without uncertainty information are described. It can be said that the above algorithm is as simple as those methods.

4. EXAMPLE OF APPLICATION: MEASUREMENT OF CONCENTRATION OF COPPER IN WATER

The GML method is applied to the data from a PT that was conducted by the Japan Society for Analytical Chemistry (JSAC) from 2014 to 2015 [15]. In this test, the concentration of copper in water was measured and reported in the unit of mg/L. It should be noted that GML analysis was not applied in the actual analysis, and the performances were evaluated through a comparison with the reference laboratory. These data are cited merely as a set of numerical examples. The following explanation is given based on the procedure described in Subsection 3.4.

Step 1: JSAC has implemented PTs for the measurement of the concentration of metals in water since 2007. Thus, information on the uncertainty caused by the measurement method has been shared by the participants, and the uncertainty has been incorporated into the reported uncertainty. Hence, it cannot be an unknown component of the uncertainty.

Step 2: The expected standard uncertainty u_{exp} was yielded as 0.0034 mg/L from the past PT data implemented from 2013 to 2014 [16]. The median of the reported relative standard uncertainties in the PT in 2013 to 2014 was 1.7 %. u_{exp} was therefore given as 1.7 % of the set concentration of copper in the PT in 2014. The set concentration was 0.200 mg/L, and u_{exp} was given as 0.0034 mg/L.

Step 3: The maximum permissible standard uncertainty of the random effect was given as 0.0010 mg/L, which is 0.3 times the expected uncertainty. Inhomogeneity of the PT items was checked in this test. The evaluated standard uncertainty between units was calculated as 0.0003 mg/L through analysis of variance. Since this value is much smaller than 0.0010 mg/L, the effect of the inhomogeneity on the PT results was negligible. In actuality, when the robust method is applied, the random effect cannot be detected.

Step 4: The minimum permissible standard uncertainty of the reported values was given as 0.0017 mg/L, which is 0.5 times the expected uncertainty. Since information on the minimum permissible standard uncertainty was not given in the actual PT, there was a laboratory that reported a standard uncertainty of less than 0.0017 mg/L. The data of that laboratory have been removed, because the data are treated only as a numerical example in the present study. Of course, it is not recommended in an actual PT that data be removed after the PT without reasonable grounds.

Step 5: The 22 reported values are shown together with their respective standard uncertainties in Table 2 and Figure 6. The reported values ranged from 0.1908 mg/L to 0.2417 mg/L. Laboratories 6 and 20 reported extremely large standard uncertainties. Unlike the case of extremely small uncertainties, these large uncertainties are not considered to impair the quality of the PT.

Step 6: The median of the standard uncertainties in Table 2 is 0.0036 mg/L, which is close to the expected standard uncertainty, 0.0034 mg/L. Since the median is larger, it can be

Table 2. Actual data reported in a PT of the concentration of copper in water conducted by the Japan Society for Analytical Chemistry from 2014 to 2015, and the evaluated values of $Q(x_i)$ and $E_{\mathrm{n}}(i)$ from the data. One set of data in which the standard uncertainty exceeded the minimum permissible standard uncertainty was removed only for the purpose of this study.

Laboratory (i)	Reported value (x_{i})	Standard uncertainty (u_{i})	$Q(x_i)$	$E_{\mathrm{n}}^{(i)}$
1	0.1908	0.0088	-162.13	-0.9
2	0.196	0.0094	-179.68	-0.5
3	0.1967	0.0033	-182.31	-1.4
4	0.1987	0.0024	-189.10	-1.4
5	0.1991	0.0050	-190.38	-0.7
6	0.202	0.11	-199.65	0.0
7	0.2025	0.0022	-201.26	-0.8
8	0.2025	0.0029	-201.26	-0.6
9	0.2043	0.0023	-205.46	-0.4
10	0.2056	0.0036	-207.32	0.0
11	0.2059	0.0033	-207.43	0.0
12	0.206	0.011	-207.42	0.0
13	0.2061	0.0044	-207.39	0.0
14	0.2070	0.0018	-206.13	0.3
15	0.2071	0.0027	-205.89	0.2
16	0.2071	0.0023	-205.89	0.3
17	0.2072	0.0018	-205.64	0.4
18	0.2116	0.0067	-185.02	0.4
19	0.2129	0.0036	-180.16	1.0
20	0.214	0.039	-176.38	0.1
21	0.216	0.014	-169.83	0.4
22	0.2417	0.0036	-127.96	4.9

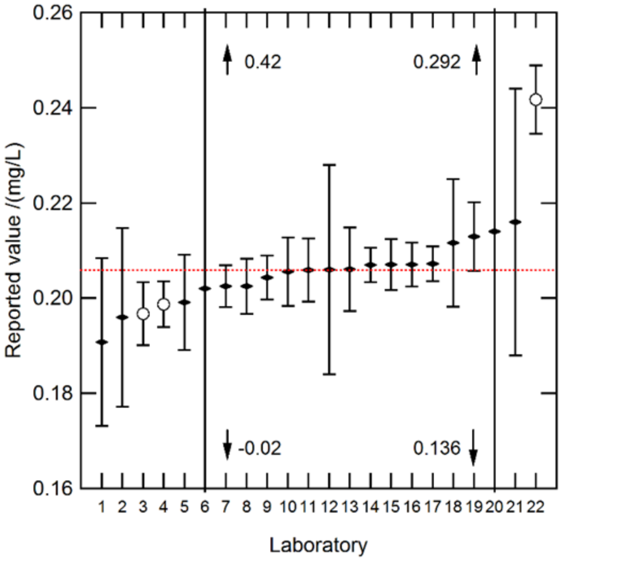


Figure 6. Actual PT data shown in Table 2. The error bars show the expanded uncertainties with a coverage factor of 2. The dotted line indicates the GML estimator, μ^{GML} = 0.2059 mg/L. The data denoted by the empty circles are the values whose performances are found to be "unsatisfactory" in the GML method.

said that a conservative evaluation was conducted to determine the maximum permissible standard uncertainty for checking the uncertainty of the random effect. On the other hand, there is a possibility such that the minimum permissible standard uncertainty of the reported value might be too small. However, as mentioned in Subsection 3.3, when the minimum permissible standard uncertainty is 0.3 times the expected standard uncertainty, the GML method works well in most cases. Since 0.0036 (mg/L) / 0.0034 (mg/L) is smaller than 0.5 / 0.3 = 1.67, this slight difference does not seem to have a significant impact on the performance evaluation.

Steps 7 to 9: The values of $Q(x_i)$ are shown in Table 2, and the minimum is given when i = 11. The value of x_{11} is therefore employed as the initial value in the repetitive computation to determine μ^{GML}. The repetitive computation in Steps 8 and 9 gives the value of μ^{GML} as 0.2059 mg/L. In this case, $\mu^{\mathrm{LML}} = \mu^{\mathrm{GML}}$, and the performances evaluated using the GML method are identical to those using the robust method.

Steps 10 and 11: The extended E_n scores are computed, and these are shown in Table 2 together with the reported values. The number of laboratories with a performance of $|E_{\mathrm{n}}^{(k)}| \leq 1$ is 19, which is more than 10. There are three laboratories whose magnitudes of E_{n} scores are larger than 1.0. It is found from Figure 6 that the values that do not contain μ^{GML} in the range of their expanded uncertainties are evaluated as "unsatisfactory". These results seem natural. We have presented a detailed discussion on the validity of the extended E_n scores in another paper [11].

For comparison, other methods were applied to this example. The robust method shows performance evaluation results identical to those obtained using the GML method. Analysis using the largest consistent subset [4] obtains E_{n} scores with magnitudes exceeding 1.0 only for Laboratories 3, 4, 19, and 22. This difference does not seem to be essential. We have provided further discussion through a comparison with other methods in another paper [12].

5. CONCLUSION

In this study, the application of maximum likelihood to the analysis of proficiency test data is discussed for cases where uncertainty information is given and a reference laboratory does not exist. To prevent serious effects from an unknown random effect and a few outliers, the following four conditions are suggested: (i) no unknown random effect other than inhomogeneity and/or instability of the PT items exists, (ii) any random effect caused by inhomogeneity and/or instability has a standard deviation of less than 0.3 × u_{exp}, (iii) the minimum permissible standard uncertainty from the laboratories is larger than 0.5 × u_{exp}, and (iv) the performances of 10 or more laboratories are consequently "satisfactory", where u_{exp} is the standard uncertainty that the participants are expected to report averagely. Based on these suggestions, a practical procedure is proposed. Moreover, the analysis method is characterized through an actual example. We believe that the analysis method proposed in this study can provide natural results with a computationally simple algorithm.

ACKNOWLEDGEMENT

This work was supported by a Grant-in-Aid for Scientific Research (No. 26870899) from the Japan Society for the Promotion of Science (JSPS).

APPENDIX: METHOD PROPOSED IN OUR PREVIOUS STUDIES

The method proposed in our previous studies [11], [12], which is referred to as the robust method in the main manuscript, is explained. It should be noted that the meanings of some symbols are different from those in the main manuscript.

In this method, the model selection is implemented through a comparison of the marginal likelihood. The statistical model with the following parameters is considered:

1. the number of data to which the common random effect is given, m (m = 0, 2, 3, …, n);
2. the identification vector for correspondence of the laboratory and the data, $\boldsymbol{v}_{\mathrm{K}} = (K(1), K(2), \ldots, K(n))^{\mathrm{T}}$ ($K(1) < K(2) < \ldots < K(m)$, $K(m+1) < K(m+2) < \ldots < K(n)$); and
3. the parameters for the priors α, β_{m+1}, β_{m+2}, …, and β_n. ($1 \leq \alpha < +\infty$, $1 \leq \beta_i$ (i = 1, 2, …, n));

where n is the number of participating laboratories.

Suppose that Laboratory $K(i)$ reports the measurement value x_i and its standard uncertainty u_i (i = 1, 2, …, n). Let $q_i = u_i^2$ for simplicity of the description. x_i is assumed to be derived from the normal distribution with the same mean of μ. On the other hand, the variances of the distribution for the reported values of Laboratories $K(i)$ (i = 1, 2, …, m) are assumed to be $q_i + \theta_{\mathrm{c}}$, where θ_{c} is the variance caused by an unknown random effect. The variances for the reported values of Laboratories $K(i)$ ($i = m$ + 1, m + 2, …, n) are assumed to be $q_i + \theta_i$, where θ_i is the additional variance caused by the unskillfulness of these laboratories. Thus, the model distributions of x_i are given as follows:

$$\begin{aligned} x_i &\sim \mathrm{N}(\mu, q_i + \theta_{\mathrm{c}}) \text{ for } i = K(1), \cdots, K(m), \\ x_i &\sim \mathrm{N}(\mu, q_i + \theta_i) \text{ for } i = K(m+1), \cdots, K(n). \end{aligned} \tag{A1}$$

Defining

$$\phi_{\mathrm{c}} = \left(\sum_{i=1}^{m} \frac{1}{q_i + \theta_{\mathrm{c}}} \right)^{-1}, \phi_i = q_i + \theta_i, \tag{A2}$$

the priors of μ, ϕ_{c}, and ϕ_i ($i = m + 1, \ldots, n$), $p(\mu)$, $p(\phi_{\mathrm{c}})$, and $p(\phi_i)$, are given as follows:

$$\begin{aligned} & p(\mu) \propto 1 \ (-\infty < \mu < +\infty), \\ & p(\phi_{\mathrm{c}}) \propto \phi_{\mathrm{c}}^{-\alpha} \ \left(\phi_{\mathrm{c}} \geq \left(\sum_{i=1}^{m} q_i^{-1} \right)^{-1} \right), \\ & p(\phi_i) \propto \phi_i^{-\beta_i} \ (\phi_i \geq q_i). \end{aligned} \tag{A3}$$

The priors of θ_{c} and θ_i, $p(\theta_{\mathrm{c}})$ and $p(\theta_i)$, are given accordingly. The hyperparameters of m, $\boldsymbol{v}_{\mathrm{K}}$, α, and β_i, are optimized to maximize the following modified marginal likelihood:

$$\Lambda = \int_{\mathrm{W}} l(\mu, \theta_{\mathrm{c}}, \boldsymbol{\theta} | \boldsymbol{x}, m, \boldsymbol{v}_K) p(\theta_{\mathrm{c}} | \alpha) \prod_{i=m+1}^{m} p(\theta_i | \beta_i) d\mu d\theta_{\mathrm{c}} d\boldsymbol{\theta}, \tag{A4}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, $\boldsymbol{\theta} = (\theta_{m+1}, \ldots, \theta_n)^{\mathrm{T}}$, and $\mathrm{W} = \{\mu, \theta_{\mathrm{c}}, \boldsymbol{\theta} | -\infty < \mu < +\infty, 0 < \theta_{\mathrm{c}} < +\infty, 0 < \theta_i < +\infty\}$. $l(\mu, \theta_{\mathrm{c}}, \boldsymbol{\theta} | \boldsymbol{x}, m, \boldsymbol{v}_{\mathrm{K}})$ is the likelihood of μ, θ_{c}, and $\boldsymbol{\theta}$ given $\boldsymbol{x}$, m, and $\boldsymbol{v}_K$. The marginal likelihoods with $m = 0$ and $m = n$ are referred to as Λ_0 and Λ_n, respectively, and employed in the main manuscript. It should be noted that $\boldsymbol{v}_{\mathrm{K}}$ is not a parameter when $m = 0$ and $m = n$.

The point is that if $m \geq 2$ is chosen as the optimized parameter, the performance evaluation should not be implemented, because $m \geq 2$ means $\theta_{\mathrm{c}} > 0$. θ_{c} is the variance of a random effect. The effect must be corrected before the performance evaluation.

Only when $m = 0$ is chosen, the performance evaluation is given. For the performance evaluation, let the posterior mean of μ with the optimized model be μ_{rob}. Several combinations of (μ, θ_1, ..., θ_n) locally maximizing the likelihood of the statistical model $x_i \sim N(\mu, q_i + \theta_i)$ ($i = 1, 2, \ldots, n$) can exist. However, the LML estimators of μ and θ_i, μ^{LML} and θ_i^{LML}, are specifically defined as the values of μ and θ_i included in the combinations of (μ, θ_1, ..., θ_n) whose value of μ is the closest to μ_{rob} among those several combinations.

Defining $\phi_i^{\mathrm{LML}} = q_i + \theta_i^{\mathrm{LML}}$, the extended E_n score for Laboratory k is proposed as follows:

$$E_n^{(k)} = \frac{x_k - \left(\sum_{i \neq k}^{n} 1/\phi_i^{\mathrm{LML}} \right)^{-1} \left(\sum_{i \neq k}^{n} x_i / \phi_i^{\mathrm{LML}} \right)}{2\sqrt{u_k^2 + \left(\sum_{i \neq k}^{n} 1/\phi_i^{\mathrm{LML}} \right)^{-1}}}, \tag{A5}$$

to check the validity of the following statistical model:

$$\begin{aligned} & x_k \sim \mathrm{N}(\mu, q_k), \\ & x_i \sim \mathrm{N}(\mu, \phi_i^{\mathrm{LML}}) \ (i = 1, \ldots, k-1, k+1, \ldots, n). \end{aligned} \tag{A6}$$

REFERENCES

[1] International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), "ISO/IEC 17025:2005: General requirements for the competence of testing and calibration laboratories", ISO, Geneva, 2005.

[2] ISO/IEC, "ISO/IEC 13528:2015: Statistical methods for use in proficiency testing by interlaboratory comparisons", ISO, Geneva, 2015.

[3] ISO/IEC, "ISO/IEC 13528:2005: Statistical methods for use in proficiency testing by interlaboratory comparisons", ISO, Geneva, 2005.

[4] M. G. Cox, Metrologia 39 (2002) pp. 589–595.

[5] M. G. Cox, Metrologia 44 (2007) pp. 187–200.

[6] R. C. Paule and J. Mandel, J. Res. Natl. Inst. Stand. Technol. 87 (1982) pp. 377–385.

[7] B. Toman, A. Possolo, Accred. Qual. Assur. Vol. 14 (2009), pp. 553–563.

[8] R. Willink, "Advanced mathematical and computational tools in metrology and testing X", World Scientific, Singapore, 2015, ISBN: 978-981-4678-61-2, pp. 78–89.

[9] S. K. Shirono, H. Tanaka, K. Ehara, Metrologia 47 (2010) pp. 444–452.

[10] R. N. Kacker, A. Forbes, R. Kessel, K.-D. Sommer, Metrologia 45 (2008) pp. 512–23.

[11] K. Shirono, H. Tanaka, M. Shiro, K. Ehara, Measurement 83 (2016) pp. 135-143.

[12] K. Shirono, H. Tanaka, M. Shiro, K. Ehara, Measurement 83 (2016) pp. 144-152.

[13] A. G. Steele, B. M. Wood, R. J. Douglas, Metrologia 38 (2001) pp. 483–8.

[14] M. J. T. Milton, M G Cox, Metrologia 40 (2003) pp. L1-L2.

[15] The Japan Society for Analytical Chemistry (JSAC), JSAC/PTP-43: Report on the proficiency test in accordance with ISO/IEC 17043, JSAC, Tokyo, 2015.

[16] JSAC, JSAC/PTP-40: Report on the proficiency test in accordance with ISO/IEC 17043, JSAC, Tokyo, 2014.