XXI IMEKO World Congress "Measurement in Research and Industry"
August 30 – September 4, 2015, Prague, Czech Republic

PROFICIENCY TEST WITH THE INFORMATION OF UNCERTAINTY: ANALYSIS WITH THE MAXIMUM LIKELIHOOD METHOD

Katsuhiro Shirono [1], *Masanori Shiro* [2], *Hideyuki Tanaka* [3], *Kensei Ehara* [4]

[1] National Metrology Institute of Japan, Tsukuba, Japan, k.shirono@aist.go.jp
[2] National Metrology Institute of Japan, Tsukuba, Japan, shiro@ni.aist.go.jp
[3] National Metrology Institute of Japan, Tsukuba, Japan, tanaka-hideyuki@aist.go.jp
[4] National Metrology Institute of Japan, Tsukuba, Japan, ehara.kensei@aist.go.jp

Abstract – The analysis methods of proficiency test (PT) data, when the uncertainty information is given, and a reference laboratory does not exist, are investigated in this study. Mainly two methods are characterized; the analyses with (i) the robust method which we developed, and (ii) the maximum likelihood estimation. Consequently, it is confirmed that the applicability latter approach is validated only when some conditions are satisfied. Therefore, the former analysis is basically recommended.

Keywords: proficiency test, ISO 13528, E_n number, measurement uncertainty, posterior predictive check

1. INTRODUCTION

The proficiency test (PT) with the interlaboratory comparison is an effective tool to assure the quality of the measurement of the calibration and testing laboratory. The participation to the PT is usually required as one of the methods to validate the measurement capability.

Usually, for the performance evaluation in the PT with the information of uncertainty, the comparison with the result of the reference laboratory is implemented using the E_n number as shown in ISO 13528 [1], which is the standard for the statistical methods for the PT tests. Given that the measurement value and its expanded uncertainty of Laboratory k are respectively x_k and U_k while those of the reference laboratory are respectively X_{ref} and U_{ref}, the E_n number for Laboratory k is defined as follows:

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

where the superscript of (k) means Laboratory k. When $|E_n^{(k)}| \leq 1$ and > 1, the performance of Laboratory k is evaluated as "satisfactory" and "unsatisfactory", respectively.

The performance evaluation for the cases where the appropriate reference laboratory does not exist has not yet been described in ISO 13528. However, in the key comparison test, which is the PT for the national metrology institutes, the comparison is basically implemented without a specific reference laboratory. Cox [2] offered a guideline of the statistical methods for the key comparison. Moreover, the analysis with the largest consistent subset (LCS) also proposed by Cox [3] has been popularly employed in the analysis. The LCS is the subset with the largest data size among the subsets whose consistencies are confirmed through the χ^2 test. Some other methods have been proposed so far [4 − 11]. It is worth to be noted that the statistical models employed in these proposals are different from each other.

We also developed the analysis method mainly focusing on the detection of an unknown random effect which impairs the quality of the PT seriously [9]. This method is robust to the outliers. Furthermore, the method is extended to be applicable in the performance evaluation. The summary of this method is given in Section 2. This method is referred to as the robust method in this paper, meaning that the method can give the robust representative value for almost all cases.

From the computational complication, this robust method may give the impression to be slightly hard to use. This does not mean that too lengthy computation time is necessary, but rather that the computation program could be complicated. The key of this method is to determine the robust representative value for the PT. Therefore, some readers consider that a simpler method like the maximum likelihood estimation can be a candidate to determine the representative value.

In this study, the conditions of the applicability of the maximum likelihood estimation are shown. Together with those, the efficacy of the robust method is confirmed. Consequently, two conditions are suggested for the application of the analysis with the maximum likelihood estimation. Therefore, the analysis with the robust method is recommended, at least from the applicability.

This paper is organized as follows: Section 2 provides the basic theory of the robust method. The comparison among the robust method, the maximum likelihood estimation, and the other methods is shown in Section 3. The brief conclusion is offered in Section 4.

2. THEORY AND PROCEDURE OF THE ROBUST METHOD

In this Section, the robust method we developed is explained. In this method, the model section (or the optimization of the parameters of the model) is implemented through the comparison of the marginal likelihood. The statistical model with the following parameters is considered:

1. The number of the data to which the common random effect is given; m ($m = 0, 2, 3, \ldots, n$) ,
2. The identification numbers for the correspondence of the laboratory and the data; $K(1), K(2), \ldots, K(n)$ ($K(1) < K(2) < \ldots < K(m), K(m+1) < K(m+2) < \ldots < K(n)$),
3. The parameters for the prior α, $\beta_{m+1}, \beta_{m+2}, \ldots$, and β_n. ($1 \le \alpha < +\infty$, $1 \le \beta_i$ ($i = 1, 2, \ldots, n$)),

where n is the number of the participant laboratories.

Suppose that Laboratory $K(i)$ reports the measurement value x_i and its standard uncertainty u_i ($i = 1, 2, \ldots, n$). Let $q_i = u_i^2$ for simplicity of the description. Since all of x_i are the measurement values for the same measurand, 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 $K(i)$ ($i = 1, 2, \ldots, m$) are assumed to be $q_i + \theta_c$, where θ_c is the variance caused by the unknown random effect. The variances for the reported values of Laboratories $K(i)$ ($i = m + 1, m + 2, \ldots, n$) are assumed to be $q_i + \theta_i$, where θ_i is the variance caused by the unskillfulness of these laboratory. Thus, the model distributions of x_i are given as follows:

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

Defining

$$\phi_c = \left(\sum_{i=1}^{m} \frac{1}{q_i + \theta_c} \right)^{-1}, \phi_i = q_i + \theta_i, \tag{2}$$

the priors of μ, ϕ_c, and ϕ_i ($i = m + 1, \ldots, n$), $p(\mu)$, $p(\phi_c)$ and $p(\phi_i)$, are given as follows:

$$\begin{aligned} p(\mu) &\propto 1 \quad (-\infty < \mu < +\infty), \\ p(\phi_c) &\propto \phi_c^{-\alpha} \quad \left(\phi_c \ge \left(\sum_{i=1}^{m} q_i^{-1} \right)^{-1} \right), \\ p(\phi_i) &\propto \phi_i^{-\beta_i} \quad (\phi_i \ge q_i). \end{aligned} \tag{3}$$

The priors of θ_c and θ_i, $p(\theta_c)$ and $p(\theta_i)$, are given accordingly. The hyper-parameters of m, $K(i)$, α and β_i are optimized to maximize the following modified marginal likelihood:

$$\Lambda = \int_W \int_{-\infty}^{+\infty} l\left(\mu, \theta_c, \boldsymbol{\theta} \middle| \boldsymbol{x}\right) p(\theta_c) \prod_{i=m+1}^{m} p(\theta_i) d\mu d\theta_c d\boldsymbol{\theta} \tag{4}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, and $\boldsymbol{\theta} = (\theta_{m+1}, \ldots, \theta_n)^T$. $l(\mu, \theta_c, \boldsymbol{\theta}|\boldsymbol{x})$ is the likelihood of m, θ_c and $\boldsymbol{\theta}$ given $\boldsymbol{x}$.

The point is that if $m \ge 2$ is chosen as the optimized parameter, the performance evaluation should not be implemented. Because $m \ge 2$ means $\theta_c > 0$. θ_c is the variance of a random effect (ex. the instability and the inhomogeneity of the measured items and the vagueness in the definition of the measurand). The effect must be corrected before the performance evaluation.

Only when $m = 0$ is chosen, the performance evaluation is given. However, in the optimized model, the additional variances are considered. In this analysis, the additional uncertainties other than the reported ones are taken in to consideration. However, a performance evaluation should serve as an index to check the validity of the reported uncertainty. Therefore, additional uncertainties must not be incorporated into the statistical model.

Thus, the following statistical model is proposed to evaluate the performance of Laboratory k:

$$\begin{aligned} x_k &\sim \mathrm{N}\left(\mu, q_k\right), \\ x_i &\sim \mathrm{N}\left(\mu, \phi_i^{\mathrm{LML}}\right) \ (i = 1, \ldots, k-1, k+1, \ldots, n), \end{aligned} \tag{5}$$

where ϕ_i^{LML} is the local maximum estimator of ϕ_i. In actuality, letting the posterior mean of μ with the optimized model as μ_{rob}, ϕ_i^{LML} is determined through the algorithm entailing the following five steps:

Step 1 $\quad \mu^{\mathrm{temp}} = \mu_{\mathrm{rob}}$

Step 2 $\quad \phi_i^{\mathrm{LML}} = \max\left[q_i, \left(x_i - \mu^{\mathrm{temp}}\right)^2 \right]$

Step 3 $\quad \mu^{\mathrm{LML}} = \left(\sum_{i=1}^{n} 1/\phi_i^{\mathrm{LML}} \right)^{-1} \left(\sum_{i=1}^{n} x_i/\phi_i^{\mathrm{LML}} \right)$

Step 4 $\quad$ If $\left| \mu^{\mathrm{LML}} - \mu^{\mathrm{temp}} \right| > \varepsilon \left(\sum_{i=1}^{n} 1/\phi_i^{\mathrm{LML}} \right)^{-1/2}$, then $\mu^{\mathrm{temp}} = \mu^{\mathrm{LML}}$ go to Step 2.

Step 5 $\quad$ For $i = 1$ to n, $\phi_i^{\mathrm{LML}} = \max\left[q_i, \left(x_i - \mu^{\mathrm{LML}}\right)^2 \right]$.

In Step 3, ε is just a calculation parameter with a small number like 10^{-3}. The calculation of Steps 2 and 5 is derived from the fact that, when μ is fixed, the maximum likelihood estimator of ϕ_i is given as $\max[q_i, (x_i - \mu)^2]$.

Based on the same idea as that in [11], we propose the index for the performance evaluation of Laboratory k as follows:

$$E_n^{(k)} = \frac{x_k - \left(\sum_{i \ne k}^{n} 1/\phi_i^{\mathrm{LML}} \right)^{-1} \left(\sum_{i \ne k}^{n} x_i/\phi_i^{\mathrm{LML}} \right)}{2\sqrt{u_k^2 + \left(\sum_{i \ne k}^{n} 1/\phi_i^{\mathrm{LML}} \right)^{-1}}}, \tag{6}$$

where $\Sigma_{i \ne k}^{n}$ denotes the summation for $i = 1, 2, \ldots, k - 1, k + 1, \ldots, n$. As well as E_n number in (1), when $|E_n^{(k)}| \le 1$ and > 1, the performance of Laboratory k is given as "satisfactory" and "unsatisfactory", respectively.

$E_n^{(k)}$ in (8) is the similar index to check the statistical model by the Bayesian posterior predictive check offered by Kacker, Forbes, Kessel, and Sommer [10]. The E_n number in

(1) can be derived based on their study, though the derivation was not directly mentioned in their study. In this sense, the E_n number proposed in the present study can be interpreted as the extended version of the E_n number for the case without a reference laboratory.

3. COMPARISON BETWEEN THE ROBUST METHOD AND THE MAXIMUM LIKELIHOOD METHOD

3.1. Procedure using the maximum likelihood estimation

Even in the procedure using the maximum likelihood estimation, the same likelihood is employed as that in Section 2. However, the parameter of m is fixed to 0. Furthermore, since the parameters of K(i) (i = 1, 2, …, n) are no longer the parameters of the likelihood when m = 0, $K(i) = i$ for i = 1, 2, …, n. Thus, the likelihood of the following model is considered:

$$x_i \sim \mathrm{N}(\mu, \phi_i) \ . \tag{7}$$

Actually, the following function $L(\mu)$ seems to provide some qualitative information about the mathematical features on the likelihood:

$$L(\mu) = \prod_{i=1}^{n} \left(2\pi\hat{\phi}_i(\mu)\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\hat{\phi}_i(\mu)}\right) \tag{6}$$

where $\hat{\phi}_i(\mu) = \max[q_i, (x_i-\mu)^2]$. The global maximum likelihood estimator of μ, μ^{GML}, matches the maximize of $L(\mu)$. The global maximum likelihood parameter of ϕ_i^{GML} is given as $\hat{\phi}_i(\mu^{\mathrm{GML}})$.

Then, the E_n numbers are computed with the following equation:

$$E_n^{(k)} = \frac{x_k - \left(\sum_{i\neq k}^{n} 1/\phi_i^{\mathrm{GML}}\right)^{-1}\left(\sum_{i\neq k}^{n} x_i/\phi_i^{\mathrm{GML}}\right)}{2\sqrt{u_k^2 + \left(\sum_{i\neq k}^{n} 1/\phi_i^{\mathrm{GML}}\right)^{-1}}}, \tag{8}$$

instead of (6). Define the global maximum likelihood estimator of μ as μ^{GML}.

3.2. Case 1: A consistent subset and some outliers

Fig. 1 shows an example of the PT data. $L(\mu)$ for the data in Fig. 1 is given as Fig. 2. This figure implies $L(\mu)$ have only a single mode at μ = 4.0. However, it is not true. Fig. 3 shows $L(\mu)$ in its logarithmic scale, and it is found that $L(\mu)$ has five modes. However, the other modes than that at μ = 4.0 is negligibly small. Thus, μ^{GML} = 4.0 is given. Actually, μ^{LML} is also 4.0 in this case.

Thus, it can be said that, when $L(\mu)$ has only a single significant mode and several negligible modes, $\mu^{\mathrm{LML}} = \mu^{\mathrm{GML}}$. In this case, μ^{LML} could be determined without the Bayesian analysis to obtain μ_{rob}. The analysis of the PT can be, hence, implemented in a very simple manner.

$E_n^{(1)}$, $E_n^{(2)}$, …, and $E_n^{(7)}$ are calculated to be – 8.6, –5.8, 0.0, 0.0, 0.0, 5.8, and 8.6. The LCS gives that $E_n^{(1)}$, $E_n^{(2)}$, …, $E_n^{(7)}$ = – 8.5, –5.7, 0.0, 0.0, 0.0, 5.7, and 8.5. Theses means the reasonable performance evaluations are given in this case with both of the robust method and the maximum likelihood method.

3.2. Case 2: A consistent subset and an extreme outlier

Fig. 4 shows a set of another virtual PT data. There seem relatively consistent data of x_1 to x_6 and an extreme outlier of x_7. Fig. 5 supports this from the perspective of the likelihood. The two large peaks can be found around μ = 4

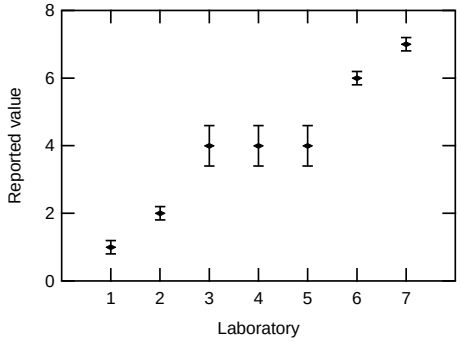

Fig. 1. Example of PT data: (x_1, x_2, …, x_7) = (1, 1, 4, 4, 4, 6, 7) and (u_1, u_2, …, u_7) = (0.03, 0.03, 0.3, 0.3, 0.3, 0.1, 0.1). Error bars show the 95 % coverage intervals.

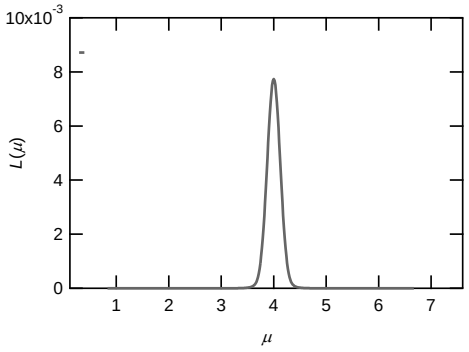

Fig. 2. Plot of $L(\mu)$ for the data in Fig. 1 in the absolute scale.

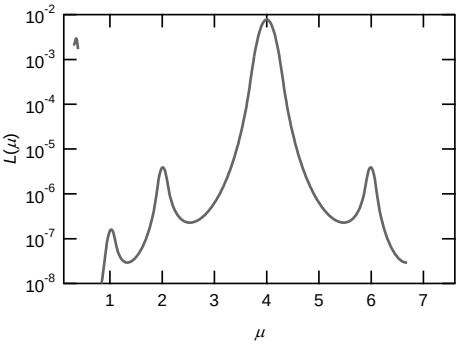

Fig. 3. Plot of $L(\mu)$ for the data in Fig. 1 in the logarithmic scale.

and 7, which match to (x_3, x_4, x_5) and x_7, respectively. It can be said from the magnitude relation of $L(4)$ and $L(7)$ that μ = 7 is more likely than μ = 4. Precisely, μ^{GML} = 7.0. On the other hand, μ^{LML} = 4.0. Thus, $\mu^{\mathrm{LML}} \neq \mu^{\mathrm{GML}}$ in this case.

With the maximum likelihood estimation, $E_n^{(1)}$, $E_n^{(2)}$, …, $E_n^{(7)}$ are given as – 3.0, – 2.5, – 1.5, -– 1.5, – 1.5, – 0.5, and 1.0. The results imply that this method is obviously influenced by the outlier of x_7 = 7.

On the other hand, With the robust method, $E_n^{(1)}$, $E_n^{(2)}$, …, $E_n^{(7)}$ = – 1.4, –0.9, 0.0, 0.0, 0.0, 0.9, and 3.0. Also, with the LCS analysis, $E_n^{(1)}$, $E_n^{(2)}$, …, $E_n^{(7)}$ = – 1.1, –0.6, 0.6, 0.6, 0.6, 1.4, and 4.5. While the quantitative trends are quite different between them, only the result of Lab. 6 is different from the qualitative perspective (i.e. "satisfactory" and "unsatisfactory"). At least, it can be said that the results with the LCS analysis is strongly affected by the reported value of x_1, whose performance is regarded as "unsatisfactory". Hence, we believe the analysis with the robust method offers the better results.

It can be concluded that, when there is only one extreme outlier, the analysis with the maximum likelihood estimation gives unnatural results.

3.3. Case 3: Data with an unknown random effect

The most typical difference between the analyses with the robust method and the maximum likelihood method can be seen at the detection of an unknown random effect. Fig. 6 shows a set of a virtual PT data. In this example, there is no consistent data. Apparently, $L(\mu)$ has the seven local maximums which corresponds to x_1 to x_7, respectively. In this case, not m = 0 but m = 7 is chosen in the robust method. Thus, the warning is emitted that the some corrections are necessary before the performance evaluation.

On the other hand, the analysis with the maximum likelihood estimator cannot give the warning principally. μ^{GML} is given as 4.0, and $E_n^{(1)}$, $E_n^{(2)}$, …, $E_n^{(7)}$ are calculated to be – 5.5, – 3.7, – 1.9, 0.0, 1.9, 3.7, and 5.5. It is technically possible that while only Lab. 3 has the adequate skill in this measurement, the other laboratories need to improve their measurement systems. However, at least, we need to consider the possibility that the quality of the PT is impaired seriously by an unknown random effect such as the instability and the inhomogeneity of the measured items and the vagueness in the definition of the measurand. Only after the inexistence of these kinds of the random effect is confirmed from the technical aspect, the calculated E_n numbers has the meaning to evaluate the performance of the laboratories.

It can be concluded that, when there is the possibility that the quality of the PT is impaired by an unknown random effect, the analysis with the maximum likelihood estimator is not recommended to be employed.

In this case, since there are no consistent subsets, the LCS analysis cannot be implemented. However, the same

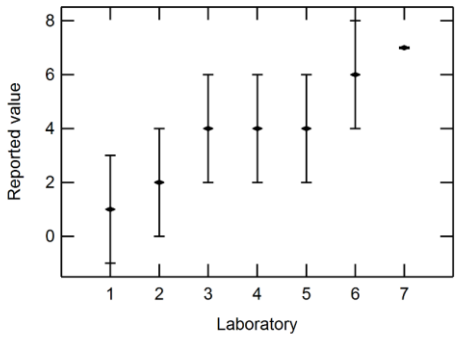

Fig. 4. Example of PT data: (x_1, x_2, …, x_7) = (1, 1, 4, 4, 4, 6, 7) and (u_1, u_2, …, u_7) = (0.03, 0.03, 0.3, 0.3, 0.3, 0.1, 0.1). Error bars show the 95 % coverage intervals.

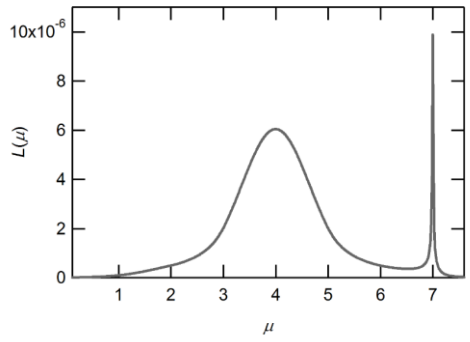

Fig. 5. Plot of $L(\mu)$ for the data in Fig. 4 in the absolute scale.

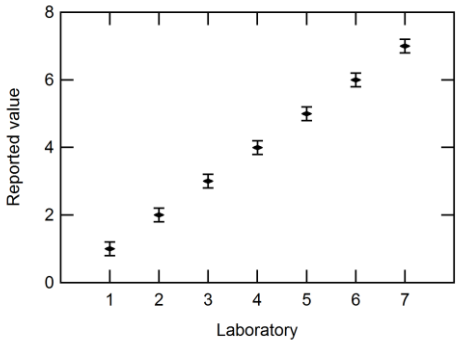

Fig. 6. Example of PT data: (x_1, x_2, …, x_7) = (1, 1, 4, 4, 4, 6, 7) and (u_1, u_2, …, u_7) = (0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2). Error bars show the 95 % coverage intervals.

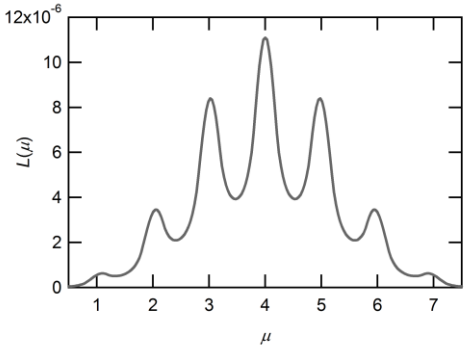

Fig. 7. Plot of $L(\mu)$ for the data in Fig. 6 in the absolute scale.

can be said to the LCS analysis.

3.4. Discussion

From the analysis of Cases 1-3, it can be said that the analysis with the maximum likelihood estimator of μ^{GML} is applicable when

(i) no extreme outliers are reported in the PT, and
(ii) the inexistence of an unknown random effect is confirmed through the technical aspect.

Otherwise, the analysis with the robust method is recommended. The second point must be investigated, because some random effects are reported even in an actual PT. On the other hand, an extreme outlier which affects the applicability of the analysis with the maximum likelihood estimation is not usually reported.

We believe the analysis with the maximum likelihood is adequately applicable in most of PTs. However, it is also confirmed that the efficacy of the robust method is not impaired in both of the cases. Thus, the analysis with the robust method is always recommended.

4. CONCLUSIONS

The analysis methods of PT data, when the uncertainty information is given, and a reference laboratory does not exist, are investigated in this study. The two conditions with which the analysis with the maximum likelihood estimation is applicable are clarified as follows: (1) No extreme outliers are reported in the PT. (2) The inexistence of an unknown random effect is confirmed through the technical aspect. Moreover, the efficacy of the analysis with the robust method is shown even in the case in which the analysis with the maximum likelihood estimation is inapplicable. Consequently, it can be said that the analysis with the robust method is always recommended, when the computational program can be properly prepared.

ACKNOWLEDGMENTS

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

REFERENCES

[1] ISO/IEC, *ISO/IEC 13528: Statistical methods for use in proficiency testing by interlaboratory comparisons*, ISO, Geneva, 2005.
[2] M. G. Cox, “The evaluation of key comparison data”, *Metrologia*, vol. 39, pp. 589–595, 2002.
[3] M. G. Cox, “The evaluation of key comparison data: determining the largest consistent subset.”, *Metrologia*, vol. 44, pp. 187–200, 2007.
[4] R. C. Paule and J. Mandel, “Consensus values and weighting factors”, *J. Res. Natl. Inst. Stand. Technol.*, vol. 87, pp. 377-385, 1982.
[5] A. G. Chunovkina and C. Elster and I. Lira and W. Wöger
[6] B. Toman, A. Possolo, “Laboratory effects models for interlaboratory comparisons”, Accred. Qual. Assur., vol. 14, pp. 553-563, 2009.
[7] R. Willink, “Models for the treatment of apparently inconstant data”, in *Advances in Mathematical and Computational Tools in Metrology and Testing X (vol.10)*, World Scientific, Singapore (2015) pp. 78-89.
[8] S K. Shirono and H. Tanaka and K. Ehara, “Bayesian statistics for determination of the reference value and degree of equivalence of inconsistent comparison data”, *Metrologia*, vol. 47, pp. 444-452, 2010.
[9] K. Shirono, M. Shiro, H. Tanaka, K. Ehara, “Theory and computation program for the novel KCRV and DOE determination method by the model selection and checking”, *Advanced Mathematical and Computational Tools in Metrology and Testing*, (presentation slides), St. Petersburg, Russia, Sept. 2014. (To be open to the public)
[10] R. N. Kacker, A. Forbes, R. Kessel, and K.-D. Sommer, “Bayesian posterior predictive *p*-value of statistical consistency in interlaboratory evaluations”, *Metrologia*, vol. 45, pp. 512–23, 2008.
[11] K. Shirono, M. Shiro, H. Tanaka, K. Ehara, “Novel reference value and DOE determination by model selection and posterior predictive checking”, in *Advances in Mathematical and Computational Tools in Metrology and Testing X (vol.10)*, World Scientific, Singapore (2015) pp. 357-366.