XX IMEKO World Congress
Metrology for Green Growth
September 9−14, 2012, Busan, Republic of Korea

VALUABLE ADVICES ON COMPARISON RULES FOR UNAMBIGUOUS EVALUATION OF MEASUREMENT RESULTS

F. Härtig, K. Kniel

Physikalisch-Technische Bundesanstalt Braunschweig und Berlin, Braunschweig, Germany, frank.haertig@ptb.de

Abstract: International comparison measurements are the backbone for validating the competence of metrology institutes among one another. An informative interpretation of the results requires, however, intensive studies of the individual reports and great expertise of the readers. This is due to the lack of a clear procedure for the assessment of such comparison measurements which leads to different values which are difficult to compare. The selection of the reference value formation is, for example, open. Likewise, an equivalence value, the normalized error value (E_n value) – which is usual for comparisons – can be calculated on the basis of the standard measurement uncertainty or on the basis of the expanded measurement uncertainty. The third degree of freedom relates to the elimination of outliers which is handled in different ways. A decision of how a participant contributes to the reference value, or whether his measurement results are regarded as comparable, may – depending on the calculation carried out – one time turn out in his favour and another time to his disadvantage. Studies of international comparison measurements from the field of length show – as examples – that the different procedures applied in comparison measurements are commonplace. In addition to the usual assessment possibilities, examples are shown which clearly show the problems. With the example of gear measurement, a proposal is presented of how comparison measurements should in future be carried out clearly and transparently.

Keywords: intercomparison, evaluation rules, precision metrology, comparability, En-value, normalized error

1. INTRODUCTION

The worldwide unification of measurements and their traceability to the International System of Units (SI) are coordinated by the highest authority in the field of metrology: the Bureau International des Poids es Mesures (BIPM) which is headquartered in Paris [1]. On 14 October 1999, 39 metrology institutes (NMI) and two international organizations signed a Mutual Recognition Arrangement (MRA) [2] with the aim of mutually recognizing measurement standards and calibration and measurement certificates and facilitating their trade relations among one another. By May 2012, the number of signatories had increased to 87. Key Comparisons (KC) of single measurands strengthen mutual confidence and provide, at the same time, information about the competence of the members. On behalf of the Comité International des Poids et Mesures (CIPM), they are organized by the Consultative Committees (CCs). The Calibration and Measurement Capabilities (CMCs) of a participant are stored in the Key Comparison Data Base (KCDB) of the BIPM and are publicly available [3]. Until 21 May 2012, 792 key comparisons have been entered there [4]. Even if the organizational realization of a KC has been determined in great detail, there are great liberties regarding the calculation of a Reference Value (RV), the selection of the measurement uncertainty for the evaluation and the exclusion of participants whose measurement results do not appear suitable. As a consequence, the results of a KC can be assessed - and thus interpreted - correctly only when the reports at hand are known in detail.

2. THE PROBLEM OF THE REFERENCE VALUE

For the calculation of a RV, different evaluations methods are available [5-10]. The most commonly used are the simple mean, the weighted mean and the median. In addition, the RV value can be calculated with the aid of other methods which will not, however, be discussed here in detail.

Simple mean

In this calculation, all measurement values enter into the RV equally weighted and measurement uncertainties are not taken into account. The simple mean is calculated in accordance with equation (1).

$$x_s = \frac{1}{n}\sum_{i=1}^{n} x_i \ ; \qquad U_s(k=2) = 2 \cdot \frac{1}{n}\sqrt{\sum_{i=1}^{n} u_i^2} \tag{1}$$

x_s simple mean
n number of measurement results
x_i measurement value of participant *i*
U_s expanded measurement uncertainty of simple mean
u_i standard measurement uncertainty of participant *i*

Laboratories with strongly deviating measurement values may "draw" the simple mean – in spite of the indication of a large measurement uncertainty – into a "wrong" direction (Figure 1).

Weighted mean

In this calculation, the measurement uncertainty associated with the measurement value is taken into account in addition. Here, the measurement values of laboratories with small measurement uncertainty contribute to the formation of the mean value with larger weight. The weighted mean value is calculated in accordance with equation (2).

The procedure requires a conscientious indication of the measurement uncertainties by the participants; otherwise, measurement uncertainties which have, for example, been estimated as too small, may lead to a falsification of the RV.

$$x_w = \sum_{i=1}^{n} \frac{x_i}{u_i^2} \cdot \frac{1}{\sum_{i=1}^{n} \frac{1}{u_i^2}} ; \qquad U_w(k=2) = 2 \cdot \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{u_i^2}}} \qquad (2)$$

x_w weighted mean

n number of measurement results

x_i measurement value of participant *i*

U_w expanded measurement uncertainty of the weighted mean

u_i measurement uncertainty of participant *i*

Median

In the case of this calculation, the measurement values are sorted by their numerical value. If the number of measurement values is odd, the RV is the measurement value at the middle position of the list. If the number of the measurement values is even, the RV is calculated from the simple mean of the two measurement values at the middle position of the list (Table 1):

Table 1 Sorted measurement values

-2.5	-2.0	-1.5	-1.5	-0.5	1.5
		x_m = -1.5			

The measurement uncertainty of the median x_m can be determined via Monte Carlo simulations. This requires a corresponding evaluation software. In the example given in Figure 1, the uncertainty of the median was calculated with an evaluation software for key comparisons [11, 12].

An important property of the median is the robustness against outliers.

Comparison

In the following, the three mean values are compared on the basis of an example. The measurement results used as a basis are shown in the list in Table 2. The data correspond to a realistic distribution from gear metrology. Figure 1 shows the measurement values with their expanded measurement uncertainties. In addition, all three measurement values have been calculated and represented. The respective uncertainties of the mean values are shown in the legend.

Table 2: Measurement results of the example

participant	measurement value in µm	expanded uncertainty in µm
A	-1.5	1.5
B	1.5	3.0
C	-2.5	0.4
D	-2.0	0.5
E	-0.5	2.0
F	-1.5	0.7

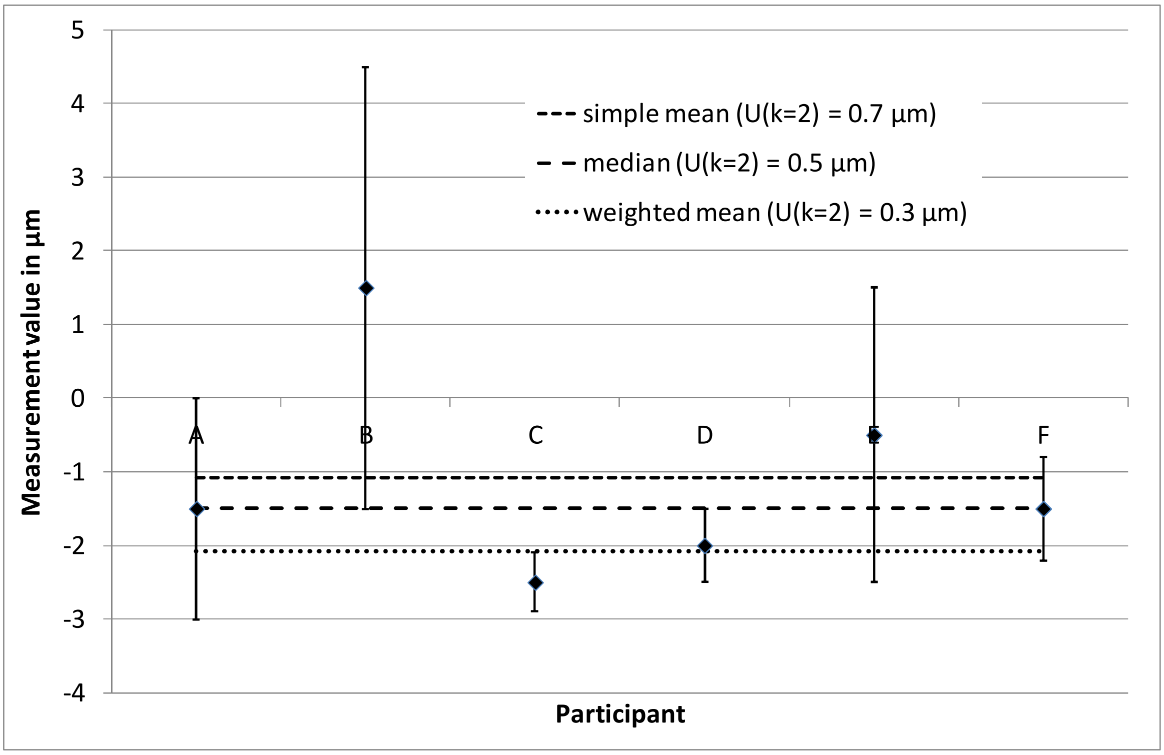

Figure 1: Example for different RV caused by the weighted mean, the simple mean and the median calculation

The three mean values differ by up to more than 1 µm. In spite of the large measurement uncertainty, the simple mean is influenced by participant B, whereas the weighted mean is drawn towards participant C due to the very low measurement uncertainty. The selection of the RV from these three possibilities will in any case affect the further evaluation, as will be shown in the following.

3. THE INCOMPARABILITY OF E_N VALUES

The agreement of a measurement result with the reference result is checked on the basis of the so-called *En* value, i.e. the normalized error. Basis are the measured values and their measurement uncertainties. If $|E_n| \leq 1$, this means that the observed measurement value x_i and the RV x_{ref} are comparable, provided that the respective measurement uncertainties are taken into account. For correlated quantities in accordance with [13], the E_n value is calculated as follows:

$$E_n = \frac{1}{k}\frac{x_i - x_{ref}}{\sqrt{u_i^2 - u_{ref}^2}} \qquad (3)$$

E_n normalized error

k coverage factor

x_i measurement value of participant *i*

x_{ref} reference value

u_i standard measurement uncertainty of participant *i*

u_{ref} standard measurement uncertainty of reference value

Standard measurement uncertainty or expanded measurement uncertainty?

According to [14] the normalized error has to be calculated on the basis of the expanded uncertainty with a confidence level of 95%. However, the problem of many comparison measurements is that the factor 1/*k* from equation (3) is omitted in the reports so that equation (4) follows. In that case, the respective standard measurement uncertainty or the expanded measurement uncertainty is used for the measurement uncertainties u_i or u_{ref}, respectively, without stating further details.

$$E_n = \frac{x_i - x_{ref}}{\sqrt{u_i^2 - u_{ref}^2}} \qquad (4)$$

To represent the effect, the E_n values for the example of Figure 1 have been calculated according to the different equations discussed in chapter 2. The values are listed in Table 3. For comparison, all three mean values have been taken as a basis and, in addition, the respective E_n value was calculated for $k = 1$ (standard measurement uncertainty) and for $k = 2$ (expanded measurement uncertainty). The latter furnishes results which differ by the amount of the coverage factor (in our example by the factor 2). For the values with grey background, $|E_n| > 1$ is valid. In the lines showing the word "undefined" no E_n value could be calculated due to a negative root (see also further below).

Table 3: E_n values based on different basis parameter

parti-cipant	simple mean		weighted mean		median	
	$\|E_n\|$ (k=1)	$\|E_n\|$ (k=2)	$\|E_n\|$ (k=1)	$\|E_n\|$ (k=2)	$\|E_n\|$ (k=1)	$\|E_n\|$ (k=2)
A	0.621	0.310	0.794	0.397	2.121	1.061
B	1.767	0.883	2.400	1.200	1.014	0.507
C	undef.	undef.	2.867	1.434	undef.	undef.
D	undef.	undef.	0.410	0.205	DIV/0	DIV/0
E	0.619	0.310	1.601	0.800	0.516	0.258
F	4.096	2.048	1.820	0.910	6.124	3.062

The free selection of the mean value and of the evaluation strategy leads to the unsatisfactory condition that the measurement results of individual participants are one time classified as comparable and the other time as not comparable with the RV.

An direct interpretation of the E_n value is, therefore, possible only when the mean value type and the coverage factor which are taken as basis are indicated.

Paradox

Large E_n values are obtained if the measurement values and the measurement uncertainties of a measurement result and of the reference result lie close together. As shown in Figure 2, this effect leads, paradoxically, to E_n values which increase with decreasing ratio of the two measurement uncertainties. This fact is shown for four differences $x_i - x_{ref}$. Using the example of the curve $x_i - x_{ref} = 1.0$, the ratios x_i to x_{ref} are visually represented at four points in an inserted diagram. Although an agreement of the values seems to be given in all four cases, taking the measurement uncertainties into account, the E_n values of the first two cases are > 1.

Measurement results which almost agree with the reference result (measurement value and measurement uncertainty) are, thus, apparently not comparable if only the E_n values are taken into account. An example is participant F in Figure 1 and Table 3 when the median is considered as the RV. Although $x_F = -1.5$ µm ± 0.7 µm and $x_{ref} = -1.5$ µm ± 0.5 µm almost agree, the E_n calculation furnish values far above 1.

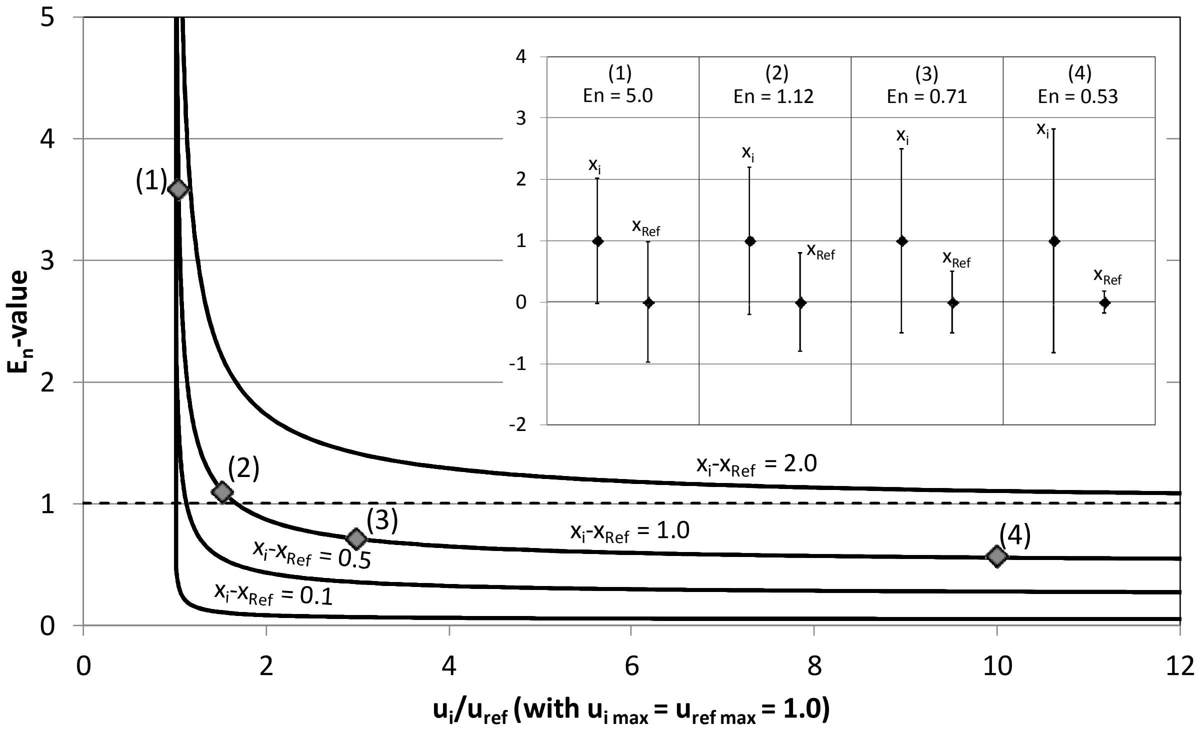

Figure 2: Example of the influence of the ratio u_i/u_{ref} in the case of an E_n value calculation.

Uncertainty of the RV always smaller than the uncertainty of the single measurements?

For the calculation of the E_n value it is probably assumed that the measurement uncertainty of the RV is always smaller than the measurement uncertainty of a single participant. That is always guarantied for the calculation of the reference uncertainty for the weighted mean but not for simple mean or median. For the two last mentioned calculation methods equation 3 or 4 do not lead to a solution. Comparison measurements, in which participants with very different measurement uncertainties take part, often do not comply with this requirement. This case can be seen in Table 3 where the E_n value based on the simple mean of participants C and D and on the median of participant C is shown. The result is that an E_n value cannot be calculated. It has not yet been determined how to proceed in such a case.

If the comparability of two measurement results is concerned and if it is tolerated that the measurement uncertainty of one participant is smaller than the measurement uncertainty of the RV, it is reasonable to consider the amount of the uncertainty difference under the root.

Improved formula for the E_n calculation

Based on the problems identified above, the equation 5 is proposed for the calculation of the E_n value. According to this equation, the E_n criterion is complied with for $E_n \leq 1$.

$$E_n(k) = \frac{1}{k} \frac{\left| x_i - x_{ref} \right|}{\sqrt{\left| u_i^2 - u_{ref}^2 \right|}} \qquad (5)$$

$$x_{ref} = x_s \text{ or } x_{ref} = x_w \text{ or } x_{ref} = x_m$$

However, the zero problem for $u_i = u_{ref}$ also remains valid for this equation.

4. THE OUTLIER PROBLEM

If the measurement results of one or several participants clearly deviate from those of the other participants, this can affect the RV significantly. Basically, the question arises of how the measurement results will be handled. A clear procedure has not been indicated here. Whereas all measurement value are taken into account for some comparisons, a formal procedure – which allows outliers to be identified and ruled out for the determination of the RV – is selected for other comparison measurements. The procedures for the identification of outliers are different and usually furnish different results. Two frequently applied methods are the use of the Birge ratio and an iteration procedure based on the consideration of the E_n values.

Here, too, a detailed study of the evaluation is indispensible to assess the RV correctly.

5. PROBLEMS OF THE REPRESENTATION OF MEASUREMENT RESULTS

The measurement results of a comparison measurement are usually represented in the form of diagrams. In addition to the measurement values and the associated measurement uncertainties, the RV is plotted. The representations do, of course, not affect the results of the interlaboratory comparison. Nevertheless, these diagrams should give a first visual information about the results obtained by the individual participants. Later, they will also provide the basis for discussions, for example in presentations. However, as the measurement uncertainties do not stand in a linear relation to one another, as is shown in the diagrams, they present – strictly speaking – a distorted picture of reality.

The measurement results are often represented as shown in Figure 3a with Table 4. The measurement values are referred to the RV which then takes the value "0" on the ordinate. The band of the associated measurement uncertainty is placed around the RV as can be seen in [15]. The measurement results are also indicated as a value with associated measurement uncertainty. In the case of this representation, the observer tends to recognize an agreement of the measurement values if the ranges of the measurement uncertainty overlap. Due to the non-linear relationships of the measurement uncertainties, this is not, however, permissible and rapidly seduce the analyst to make a false statement, as in Figure 3a in the case of participant 2. In spite of apparent agreements, the E_n criterion is not complied with.

If one selects, however, the representation shown in Figure 3b, visually meaningful and analyzable results are obtained in a simple way. In this representation, the measurement uncertainties are summarized quadratically to form a combined measurement uncertainty. This combined measurement uncertainty is then assigned to the measurement value as an error bar. It is easy to recognize that only the error bars of the measurement values, which meet the E_n criterion, cross the value of the RV.

Table 4: Measurement results and their evaluation

partici- pant	**measurement result $x_i \pm U_i$ in µm**	**reference result $x_{ref} \pm U_{ref}$ in µm**	**combined expanded measurement uncertaintyU_i* in µm**	**$\|E_n\|$**
1	0.50 ± 1.00	0.00 ± 0.58	0.65	0.61
2	1.00 ± 1.00	0.00 ± 0.58	0.65	1.22
3	-1.50 ± 1.00	0.00 ± 0.58	0.65	1.84

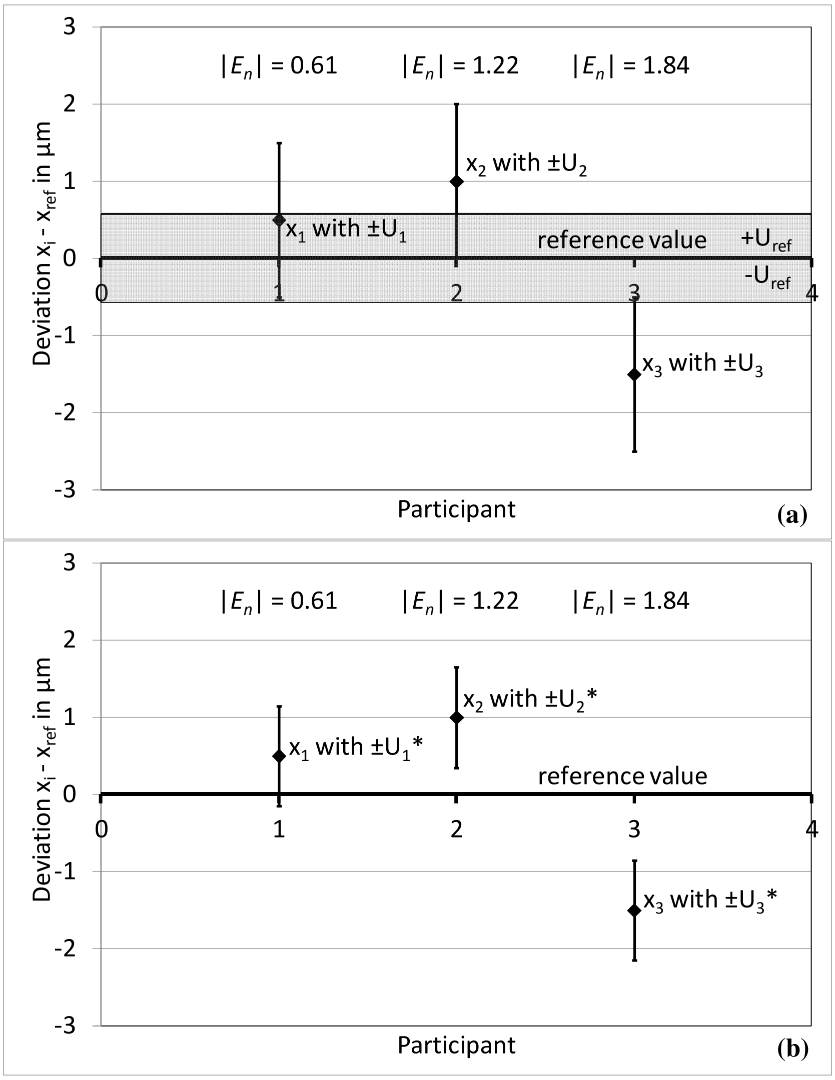

Figure 3: Representation of measurement results of a comparison measurement. (a) Separate plotting of the measurement uncertainties U_i and U_{ref}, no visual interpretation of the comparability of the measurement results possible, (b) Plotting of the combined measurement uncertainty U_i* above the measurement value, visual interpretation of the comparability of the measurement result possible

6. PROCEDURE FOR COMPARISON MEASUREMENTS IN THE FIELD OF GEAR METROLOGY

In the following, the most important steps are briefly shown – as an example – for the realization of an interlaboratory comparison:

1. Determination of the participants
2. Determination of the material standards, of the detailed measurement conditions such as, for example, cleaning, clamping, probe, determination of the measurands and their evaluation instructions
3. Determination of the return values: measurement values and the expanded measurement uncertainties, including the indication of the coverage factor
4. Calibration of the measurement standards by the pilot institute
5. Calibration of the measurement standards by the participants. Possibly planning of regular recalibrations by the pilot institute to recognize changes of the standard at an early stage.
6. Recalibration of the measurement standards by the pilot institute
7. Determination of the reference value (RV) by calculation of the weighted mean (equation 2)

$$x_{ref} = \sum_{i=1}^{n} \frac{x_i}{u_i^2} \cdot \frac{1}{\sum_{i=1}^{n} \frac{1}{u_i^2}} \; ; \qquad U_{ref}(k=2) = 2 \cdot \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{u_i^2}}}$$

x_{ref} reference value (weighted mean)

n number of measurement results

x_i measurement result of participant *i*

U_{ref} expanded measurement uncertainty of the weighted mean

u_i measurement uncertainty of participant *i*

8. Calculation of the E_n value for the coverage factor *k*=2 on the basis of the indicated expanded measurement uncertainty (equation 5)

$$E_n(k) = \frac{1}{k} \frac{\left|x_i - x_{ref}\right|}{\sqrt{\left|u_i^2 - u_{ref}^2\right|}}$$

9. First information of the participants about measurement results, giving them the possibility of controlling them again. Information about possible inconsistencies by the pilot institute is possible.
10. No outlier elimination, all participants are, as a matter of principle, taken into account in the RV. The participants may, however, withdraw their results voluntarily.
11. New assessment and report of the results

6. SUMMARY

Today, the calculation of a RV used for international comparison measurements as well as the calculation of the *En* criterion, which is used to evaluate the competence of a participant, are carried out almost arbitrarily within the scope presented. The interpretation of the different comparison measurements are, therefore, possible only if the associated reports have been intensively studied. In the preparatory phase of an interlaboratory comparison, the freedom to select the evaluation procedure leads, however, also to detailed discussions, because the participants are interested in selecting a procedure which allows them to obtain favourable results in the comparison. For the field of gear measurement, a proposal for an unequivocal procedure for the performance of comparison measurements has been made. For metrology in general it would be desirable if the BIPM would also specify unequivocal rules for the assessment of comparison measurements. First activities are in progress [16].

5. REFERENCES

[1] BIPM Bureau International des Poids es Mesures. http://www.bipm.org/ (access 2012-05-31)

[2] BIPM: CIPM Mutual Recognition Arrangement. http://www.bipm.org/en/cipm-mra (access 2012-05-31)

[3] BIPM: The BIPM key comparison database. http://kcdb.bipm.org/ (access 2012-05-31)

[4] BIPM: http://www.bipm.org/utils/common/pdf/Participation_in_SCs.pdf; (access 2012-05-12)

[5] Key Comparison CCL-K1: Calibration of gauge blocks by interferometry. Final report, 2000

[6] EUROMET Key Comparison EUROMET.L-K2: Calibration of long gauge blocks. Final report, 2006

[7] Key Comparison CCL-K3: Calibration of angle standards. Final report, 2007

[8] Key Comparison CCL-K4: The Calibration of internal and external diameter standards. Final report, 2007

[9] Key Comparison CCL-K5: CMM 1D: Step gauge and ball bars. Final report V3.2, 2009

[10] Key Comparison CCL-KC-6: Calibration of coordinate measuring machine (CMM) two-dimensional (2-D) artifacts (ball plates & bore plates). Final report, 2008

[11] Douglas R., Steele A.: En Toolkit - for evaluating key Camparisons and KCRV in Excel. National research Council Canada, Institute for national measurement standards

[12] Decker J. E., Lewis A. J., Cox M. G., Steele A. G., Douglas R. J.: Evaluating results of international comparisons: Worked example of CCL-K2 comparisons of long gauge block calibration; XVIII IMEKO World Congress; Metrology for a sustainable development; September 17-22; 2006, Rio de Janeiro, Brazil

[13] Wöger W.: Remarks on the En-criterion used in measurement comparisons; PTB-Mitteilungen 109, 1/99

[14] BIPM Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes Paris, 14 October 1999. http://www.bipm.org/en/cipm-mra/mra_online.html (access 2012-05-30)

[15] Cox M.: The evaluation of key comparison data: determining the largest consistent subset; Metologiea 44 (2007) 187-200

[16] Lewis A.: Guide to preparation of Key Comparison Reports in Dimensional Metrology; CCL/WG-MRA/GD-2, 2012