On the definition of "significant unresolved deviation" in the MRA frame

Franco Pavese, IMGC-CNR, Torino, Italy

1. Introduction

The MRA prescribes that the international degree of equivalence and its uncertainty is obtained from the results of key comparisons (KC).[1] From these results, the international acceptance of the services provided by the NMI's, the CMC's, can be established.

There are two links between the uncertainty that can be accepted for the (top quality) CMC's and the results published in the Appendix B of the BIPM database for MRA:

a) the uncertainty of the CMC's must be "*consistent with the results given in Appendix B, derived from the key comparisons*". The meaning of consistent has been specified more in details by the JCRB.

b) should "*, as a result of a key comparison, a significant unresolved deviation* [SUD] *from the key comparison reference value*" occur, "*the existence of this deviation is noted in Appendix C. In this case, the institute has the choice of either withdrawing from Appendix C one or more of the relevant calibration and measurement services or increasing the corresponding uncertainties given in Appendix C.*".[2] No general definition of SUD exists in the MRA nor from the JCBR. For the temperature field, on request of the President of the CCT, the JCRB said that the CIPM should decide and the CIPM assigned the CCT the competence and the capacity to provide a definition of SUD.

2. Meaning of international acceptance of CMC's

The international acceptance of CMC's is not supposed to replace the statements of its own of each NMI concerning the uncertainties of their own services, nor the existing or future bilateral or multilateral agreements between NMI's with a similar contents.

However, the value and the authority of this acceptance under the MRA is obviously to be considered of higher rank, at least for the Countries that signed the Convention du Mètre and the MRA Protocol.

The agreement on the CMC's is obtained with a specific procedure, described in the Protocol, which also is of higher rank with respect to the self-declarations.

In fact, in addition to providing an uncertainty budget self-declared by each NMI (possibly taking into account the ongoing scientific debate on the subject, and/or the requirements set forth by the inter-comparison protocols, including the KC's), the MRA procedure also requires, for the statement of the CMC uncertainties, to demonstrate the uncertainties by taking into account the *results* of the KC's, in particular the differences found between NMI's.

Consequently, the MRA does not prevent any NMI from continuing to declare also its own uncertainty for its own services, as before, as a self-declaration.

However, in order to obtain the international acceptance of the same services under the MRA, each NMI must accept and declare the uncertainties which are stemming from the MRA procedure, i.e. from the international degree of equivalence (DoE), demonstrated through the KC results.

It is up to each NMI customer to decide whether to trust more –or to find more convenient– and use one uncertainty declaration instead of the other, each declaration having, of course, a different meaning.

[1] Organised by the CIPM Consultatif Committees. Also RMO and others levels of KC are allowed.

[2] This definition requires a KCRV be defined. For the case it is not, see [1] Doc.CCT/03-04.

A major point, which can arise from above, is that the MRA "accreditation" may result in an uncertainty higher than the self-declared one.

The reasons why this can happen are threefold:

i) contribution to the uncertainty of the KC itself;
ii) detection of type-B errors in NMI's;
iii) "spot accidents". [3]

In case i) and partially ii), a KC can happen to be unable to allow for the smallest values self-declared for the NMI uncertainties, i.e., can be unable to represent a check and a validation of these values. Case i) is a "defect" –some times unavoidable– of the KC; case ii) is the most valuable information one can get from a KC.

In case iii), there was, of course, an unfortunate occurrence.

However, there are no corrective actions possible to avoid possible penalties to NMI's with the smallest uncertainties, the aim being, as in the case of MRA, to spot the knowledge available *at a certain time*, for the purpose of establishing the international equivalence: if the demonstrated equivalence is affected by ambiguities in this knowledge –as represented by the KC results, it has to be accepted anyway (or the NMI withdraws from the KC),being that the to-date state of the art, until new knowledge becomes available in the future.

3. About the definition of SUD

The definition of differences between NMI's and their statistical significance is obviously a very delicate issue and, for this reason, should be treated strictly as a scientific issue –as opposed to a "political" one.

The starting point is that this differences can only be evaluated from the results of inter-comparisons. In the MRA frame, they are KC's, in particular CC KC's, whose results are:

1) the degree of equivalence (either, 1a) "absolute" [4] or, 1b) "mutual" [5]);
2) the uncertainty associated to 1), which the MRA prescribes be evaluated at the 95 % level of confidence. This corresponds to $k = 2$ when the underlying pdf is Gaussian, as in Section 3.1.

Except the data 1b), the other require the definition of a representative value of the KC, called KCRV. The MRA does not provide any prescription about an uncertainty to be associated to it (which, consequently, can also be zero: see [2] for a discussion of the latter case [6]) .

In general, the SUD definition involves having: i) two representative values y_1 and y_2, ii) the uncertainties associated to each of them, u_1 and u_2; iii) a criterium to establish the statistical significance of the difference $(y_1 - y_2)$.[7]

3.1 KCRV defined

If normality can be assumed for the pdf's of all participant's data, the most straightforward and sound method for iii), in agreement with mainstream GUM and according to the procedure recently published by the BIPM Director's Advisory Group on Uncertainty [3][8], is the χ^2 test. Whatever is the summary statistics used to compute y and the associated u, then

[3] I.e., should the comparison repeated, they would not occur anymore.

[4] I.e., of each NMI with respect to the KCRV.

[5] I.e., between pairs of NMI's.

[6] [2] Doc. CCT/03-05: F.Pavese and P.Ciarlini, in press on PTB Berichte (2003),

[7] The direct use of the pdf's does not change this scheme.

[8] [3] M.G.Cox, *Metrologia*, 2002, **39**, 589-595.

“ *3. Apply a chi-squared test to carry out an overall consistency check of the results obtained:*

(a) *Form the observed chi-squared value in Section 1,*

$$\chi^2_{\text{obs}} = \frac{(x_1 - y)^2}{u^2(x_1)} + \cdots + \frac{(x_N - y)^2}{u^2(x_N)}.$$

(b) *Assign the degrees of freedom*

$$\nu = N - 1.$$

(c) *Regard the consistency check as failing if*

$$\Pr\{\chi^2(\nu) > \chi^2_{\text{obs}}\} < 0.05.$$

Note 1. Pr denotes “probability of”.
Note 2. This test assumes normality, …” [3]

The test threshold in (c) above is set at 0.05, since “the uncertainty of … deviation at the 95 % level of confidence”, as done for all temperature CC KC.

Apart from the authoritative source of this advice, there is no statistical reason to use a different methods for testing the SUD, when the hypotheses hold.

In past discussions, the use of k = 3 for testing SUD has been proposed (while, at the same time, using k = 2 for the extended uncertainty of the DoE). The argument was that, approaching N = 20 participants, there is an increasing probability of a “false positive” SUD (1 over 20 cases).

It looks like that this interpretation is using in an incorrect sense the meaning of 95% level of confidence. In fact, the reason why a 1 over 20 probability to have a SUD applies comes because the chosen uncertainty for the DoE is “too *narrow*”, not because the test threshold is too narrow: if one takes k = 1, should there be a 33% probability of “false positive” ? Obviously not, just the uncertainty is much too narrow.

In all instances, any statistical justification sems possible for taking a different value of k for the DoE uncertainty and for the SUD test: *SUD’s come just from the decisions taken for DoE’s*.

An example of a method that combines deviations and uncertainties is the use of the QDE [9]. Should it be used to quantify the DoE uncertainty, it seems to me that it would, by definition, suppress the possibility to have any SUD –being the QDE already taking into account also the deviations– *but just at the cost of an increase of the DoE uncertainty, when deviations are present*, with respect to the one computed with simply k = 2. Also consultation of Doc.CCT/01-41 is useful in this respect.

3.2 KCRV not defined

One must assume first that, when in a KC no reference value is defined, there were very good reasons to avoid it and, consequently, it seems incorrect to introduce, for the purpose of the SUD test only, the approach that was decided *not* to use for the equivalence (the one in [3] or any other resorting to a summary random variable of the set $\{y\}$). The additional risk is that such a KCRV and its uncertainty become purely “consensus values” –i.e. “political” decisions.

SUD’s in *mutual* equivalence, automatically, are also such for the corresponding absolute equivalence –at least when u_{KCRV} = 0, as in CCT K2 and K4. Their detection, in principle can be

[9] Steele A. G., Hill K. D., Douglas R. J., *Metrologia* 2002, **39**, 269-278.

done similarly to the case 3.1, not involving the KCRV definition, however the corrective action called by the MRA need to understand how to increase the uncertainty of each NMI of the pair.

One way to resort to a representative value x_{ref} without incurring in the well know difficulties of choosing a summary statistics, especially when the KC pdf is not unimodal, is introduced in Doc.CCT/03-05: the use of the *expected value of the mixture density* function representing the KC.

Concerning the uncertainty to associate to it, clearly the 95%CI of the mixture represents the uncertainty in the actual knowledge of that fixed point, at that confidence level; on the other hand, it is not an accurate representation of the (average ?) NMI's measurement capability. One solution might again be to take $u_{\text{ref}} = 0$, as done for CCT K2 and K4, resulting in a slightly more stringent test threshold for absolute SUD's.

Another road possibly to follow is to resort to method indicated for the testing Laboratories, such as in ISO5725.

In the testing frame, it is preferred the use of a tolerance criterium (proficiency test limits, use of Z-scores). It may be the best criterium also for the test of SUD's. A tolerance-type decision can be a more appropriate frame for it: essentially it is this type of decision (yes/no) that is required for the SUD definition.