UPDATE TO PROPOSAL FOR KCRV & DEGREE OF EQUIVALENCE FOR GTRF KEY COMPARISONS

J. Randa, NIST, February 2005

INTRODUCTION

A few years ago, I proposed a method for identifying outliers and computing the Key Comparison Reference Value (KCRV) in Key Comparisons [1]. Since then, several GT-RF Key Comparisons have used some or all of those suggestions, and therefore we have gained some experience with them. Also since that time, the BIPM Director's Advisory Group on Uncertainties has issued recommendations for the evaluation of key comparison data [2,3]. The recommendations in [2,3] are not mandatory, but they represent the considered opinion of a group of experts and should obviously be given serious consideration. The present document revisits the issue of outliers and KCRV computation: it presents a summary of the recommendations of the Director's Advisory Group; it updates [1]; and it comments on both.

DIRECTOR'S ADVISORY GROUP RECOMMENDATIONS

What follows is a short summary of (some of) the recommendations of the BIPM Director's Advisory Group on Uncertainties [3]. Anyone preparing the protocol for a key comparison that may use this approach should refer to [3] for the full recommendations and detailed, step-by-step procedures. Two procedures are detailed, Procedure A and Procedure B. (Note that [2] explicitly states "The guidelines carry no formal status and thus are in no sense mandatory.")

Assuming that each participating lab provides a measurement result x_i and an associated standard uncertainty $u(x_i)$, the preferred procedure (Procedure A) is to first calculate the weighted mean

$$y = \left(\sum_1^N \frac{x_i}{u^2(x_i)} \right) \times \left(\sum_1^N \frac{1}{u^2(x_i)} \right)^{-1} \tag{1}$$

and the estimate of its uncertainty

$$u(y) = \left(\sum_1^N \frac{1}{u^2(x_i)} \right)^{-1/2} . \tag{2}$$

A chi-squared test (detailed in [3]) is then performed to check that the data are consistent. If the chi-squared consistency check does not fail, then y as given by eq. (1) is taken as the KCRV, $x_{ref} = y$, and $u(y)$ from eq. (2) is the standard uncertainty in x_{ref}. One then proceeds to compute degrees of equivalence and the associated uncertainties in a standard manner.

The advantage of this procedure is that it uses the weighted mean, which is the most defensible choice scientifically (if all submitted uncertainties are valid), and it protects against the problem of unrealistically small uncertainties by means of the

consistency check. If the consistency check fails, an attempt is made to identify the cause of the failure. Laboratories with degrees of equivalence d_i that differ from zero by more than twice the standard uncertainty in d_i are asked to check and correct their measurements (assuming Draft A has not yet been circulated). If they have no corrections, they are asked if they want to withdraw. If one or more labs change their results or withdraw, the consistency check is applied to the new set of results. If the test still fails, Procedure B is applied.

Procedure B is based on Monte Carlo simulations of the median "(or some other suitable estimator)" generated from distributions representing the results from each lab. A very large number of simulated data sets are generated, the median of each is computed, and then the KCRV is computed by taking the mean of the medians of all the simulated data sets. Use of the median reduces the sensitivity to outliers, and the average over a large number of simulated results should or could reduce the sensitivity to just one result that occurs for the median method for small numbers of participants. For present purposes, we will not treat Procedure B in detail. Anyone interested should consult [3].

UPDATE ON GT-RF/2000-12

At the time [1] was proposed, there was an immediate need for a quantitative method for identifying and dealing with possible outliers in key comparisons. Key comparisons were being approved and conducted, and a prescription was needed for computing the KCRV and degrees of equivalence, even in the presence of possibly aberrant results from one or more participants. The principal point of [1] was to use the median of absolute deviations (MAD) as a quantitative, relatively robust way to identify outliers. It was recognized that the proposed procedure was rather *ad hoc* and far from perfect, but there were not many alternatives available, and several comparisons adopted and used the suggestions of [1]. Now that we have some experience with the procedure, we can use the lessons learned to modify it and point out some limitations, all related to the case of small number of participants.

The first point is that eq. (1) of [1], which relates the sample standard deviation to the MAD,

$$\sigma \approx S(MAD) = 1.4826 \; median_j \left\{ \left| Y_j - Y_{med} \right| \right\}, \qquad (3)$$

is valid for infinite sample size. For finite sample size, a multiplier κ other than 1.4826 should be used. I do not know an exact form for κ, but from simulations one obtains the results in Table 1. Also, eqs. (6) – (8) in [1], use a coverage factor $k = 2$ for the expanded uncertainties in the degrees of equivalence. This is appropriate for very large effective degrees of freedom, but for finite degrees of freedom, the appropriate coverage factor should be determined in the manner described in Annex G of the GUM [4].

The second point relates to the expression for the uncertainty in the KCRV when the unweighted mean is used to compute the KCRV. The suggestion in [1] was that the uncertainty in this case should be computed from the distribution of the measurement results for the quantity measured, i.e., the standard deviation of the mean of the sample

$$u^2(y) = \frac{1}{N(N-1)} \sum_1^N (y_i - y)^2 , \qquad (4)$$

where y is the KCRV. This was recommended as a way to avoid using the uncertainties submitted by the participating labs. It is now clear that this was a bad idea, particularly for small samples. For small samples, it can happen that eq. (4) leads to very small uncertainties due to "accidental" very close agreement of the results y_i . It is much better to use the individual uncertainties $u(y_i)$ to compute u(y),

$$u^2(y) = \frac{1}{N^2} \sum_1^N u^2(y_i) . \qquad (5)$$

The worry that one lab might have undue influence because of its very small uncertainty is unfounded because eq. (5) is not very sensitive to one very small value for $u(y_i)$.

The final lesson learned is that for very small sample sizes, the MAD procedure can produce anomalously small values for $S(MAD)$ and the estimated standard deviation, due to random clustering. This can lead to very questionable outlier identification. For example, if the MAD criterion is applied to the data set {505, 501, 500, 500, 498}, the point at 5.05 would be classified as an outlier—even though the uncertainty on each point might be 1 % or 2 %. (The MAD makes no use of the uncertainties of the individual points.) Such a situation should not occur often, nor should it occur for large data sets, but it does occur for small sample sizes, and it is troublesome when it occurs. In such a case, one hopes that the participants would agree to include the "outlier" in the computation of the KCRV. (Note that even if the point is excluded from the calculation of the KCRV, its degree of equivalence can still be consistent with 0.)

COMMENTS AND SUMMARY

The Director's Advisory Group's Procedure A uses a weighted mean for the KCRV, with a chi-squared test to check that the data (including uncertainties) are consistent. If the chi-squared test fails, Procedure B is used. Procedure B looks reasonable, but it is less transparent and more computationally intensive than Procedure A. If there is resistance to using the weighted mean, it appears that this approach could also be used if the KCRV were computed from an unweighted mean. In that case the chi-squared would not attain its minimum value (the minimum occurs for the weighted mean), but it would still test whether the data were consistent with the computed KCRV.

Two modifications of the MAD method were given. The multiplier 1.4826 in eq. (3) should be replaced by the value of κ from Table 1 for finite samples, and the correct coverage factor for finite degrees of freedom should be used in obtaining the expanded uncertainty. The uncertainty in the KCRV should be computed from the participants' uncertainties, as in eq. (5). In addition, a problem that can arise in very small samples was noted.

References:

[1] J. Randa, "Proposal for KCRV and Degree of Equivalence for GTRF Key Comparisons," GT-RF/2000-12, August 2000.

[2] M.G. Cox, "The evaluation of key comparison data: An introduction," Metrologia, **39**, 587 – 588, 2002.

[3] M.G. Cox, "The evaluation of key comparison data," Metrologia, **39** 589 – 595, 2002.

[4] Guide to the Expression of Uncertainty in Measurement, ISO, 1993.

Table 1. Multiplier κ for eq. (3).

N	κ
2	1.773
3	2.206
4	2.019
5	1.800
6	1.764
7	1.686
8	1.671
9	1.633
10	1.626
11	1.602
12	1.596
13	1.581
14	1.577
15	1.566
20	1.544
25	1.530
50	1.507
100	1.494
1000	1.484
2000	1.483