INSTITUTE OF PHYSICS PUBLISHING METROLOGIA
Metrologia **42** (2005) L13–L14 doi:10.1088/0026-1394/42/1/L04

SHORT COMMUNICATION

Response to comments on 'Statistical analysis of CIPM key comparisons based on the ISO *Guide*'

R N Kacker, R U Datla and A C Parr

National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA

Received 6 January 2005
Published 9 February 2005
Online at stacks.iop.org/Met/42/L13

Abstract
The authors respond to the comments on their paper made in the preceding article.

We are very pleased that Franco Pavese's and our viewpoints on the statistical analysis and interpretation of CIPM key comparison are very similar. Franco Pavese has commented on two papers [1, 2]. As noted by him, some of our points have previously been made by him and his collaborators in various papers and presentations. Franco Pavese presents his comments in three sections: (i) key comparison data model, (ii) the value of the measurand and the mixture distribution and (iii) kinds of key comparisons. We respond to these sections.

1. Key comparison data model

Pavese is correct; the systematic laboratory effects model proposed in [1, 2] is a method to account for the uncertainty that arises from an unknown bias (in an uncorrected combined result such as a weighted mean or an arithmetic mean of the laboratory results) and there is no disadvantage to adopting it for the statistical analysis of key comparison data.

We believe that the purpose of CIPM key comparisons is not proficiency testing of the participants. Our interpretation of the CIPM/BIPM publications is that the participants of a CIPM key comparison are specific world-class laboratories that are believed to be competent for the measurement techniques being investigated. The competency of a candidate laboratory, when in question, should be investigated before the key comparison. Once a key comparison has begun, all participants are to be treated equally. Despite the competency of participants, the results and/or statements of uncertainty submitted by some may be unreasonable in view of the previous or other measurements and the well understood uncertainties associated with the measurement technique. Therefore, some Consultative Committees (CCs) may choose to screen (and possibly adjust) the data published in Draft-A before using them to determine the key comparison reference value and its associated uncertainty. As suggested in [1], a useful statistical method for flagging discrepant results and uncertainties is the ASTM documentary standard E691-1999 [3].

The key comparison reference value (KCRV) does not have a single meaning that applies to all key comparisons. Also, some key comparisons are not appropriate for establishing the KCRV. For example, the CC on temperature did not establish the KCRV from the key comparison CCT-K3 [4]. Generally speaking, the purpose of a CIPM key comparison is to establish—when appropriate—the KCRV, the degrees of equivalence and their associated uncertainties on the basis of the data provided by the participants. The systematic effects model is useful for the statistical analysis of data from such key comparisons.

The International Organization for Standardization *Guide to the Expression of Uncertainty in Measurement* (ISO-GUM) is not consistent with the sampling theory (frequentist) statistics [5]. Therefore, sampling theory models are not appropriate for the statistical analysis of key comparisons. The systematic laboratory effects model is based on the ISO-GUM.

2. The value of the measurand and the mixture distribution

We were not aware; now, we are aware that Pavese and his collaborators had proposed in a 2001 conference presentation [6] the use of a mixture probability distribution for the statistical analysis of temperature key comparisons. He and his collaborators discussed the bimodal form of the mixture distribution in [7].

The ISO-GUM recommends propagation of uncertainties using a linear approximation of the measurement equation. In the ISO-GUM, the inputs to a measurement equation are the expected values and standard deviations of state-of-knowledge probability distributions for the input variables [8, section 4.1.6]. The output is the expected value and standard deviation of a state-of-knowledge probability distribution for the output variable. The result and standard uncertainty for the value of the measurand determined by propagating uncertainties according to the ISO-GUM are interpreted as the expected value and standard deviation (both approximate) of a state-of-knowledge probability distribution for the value of the measurand. The result and standard uncertainty for the value of the measurand so determined are fit for all probability distributions that have the specified expected values and standard deviations for the input variables. The ISO-GUM assumes less and hence it is more robust than every approach that combines or propagates probability distributions. In particular, the ISO-GUM is more robust than the use of a mixture probability distribution. In [1], we discussed the use of the expected value and standard deviation only of a mixture distribution. Our proposal in [1] to use only the expected value and standard deviation of a mixture distribution is as robust as the ISO-GUM. This of course requires the assumption that the result and standard uncertainty for the value of the measurand determined by propagating uncertainties are meaningful and appropriate for the particular application.

3. Kinds of key comparisons

Pavese and his collaborators [7, 9] classified key comparisons into two classes based on the nature of the measurand. The two classes are as follows: artefact intercomparisons (class 1) and physical-state realization intercomparisons (class 2). In [1], we classified key comparisons into two kinds based on the nature of the laboratory results. The two kinds are as follows: the laboratory results are direct measurements of a common measurand of a stable value during the comparison (first kind) and the laboratory results are not direct measurements of a stable measurand (second kind). Pavese has raised several issues concerning the relationship between his and our classifications and their consequences. These issues need further investigation.

References

[1] Kacker R N, Datla R U and Parr A C 2004 Statistical analysis of CIPM key comparisons based on the ISO Guide *Metrologia* **41** 340–52

[2] Kacker R N, Datla R U and Parr A C 2003 Statistical interpretation of key comparison reference value and degrees of equivalence *J. Res. Natl Inst. Stand. Technol.* **108** 439–46

[3] 1999 *Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method* ASTM standard E691 http://www.astm.org

[4] Mangum B W, Strouse G F and Guthrie W F 2002 CCT-K3: Key comparison of realizations of the ITS-90 over the range 83.8058 K to 933.473 K *Technical Note 1450* (National Institute of Standard and Technology)

[5] Kacker R N and Jones A T 2003 On use of Bayesian statistics to make the *Guide to the Expression of Uncertainty in Measurement* consistent *Metrologia* **40** 235–48

[6] Ciarlini P, Pavese F and Regoliosi G 2002 *Algorithms for Approximation IV (Huddersfield Conference, July 2001)* ed I J Anderson *et al* (University of Huddersfield, Huddersfield, UK) pp 138–45

[7] Ciarlini P, Cox M, Pavese F and Regoliosi G 2004 The use of a mixture of probability distributions in temperature interlaboratory comparisons *Metrologia* **41** 116–21

[8] 1995 *Guide to the Expression of Uncertainty in Measurement* 2nd edn (Geneva: International Organization for Standardization) ISBN 92-67-10188-9

[9] Pavese F and Ciarlini P 2003 Classes of inter-comparisons and the case of temperature standards *PTB Berichte* PTB-IT-10, 63–76