

Recent recommendations of the Consultative Committee for Length (CCL) regarding strategies for evaluating key comparison data

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2006 Metrologia 43 L51

(<http://iopscience.iop.org/0026-1394/43/6/N06>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 38.107.179.214

The article was downloaded on 21/02/2012 at 01:23

Please note that [terms and conditions apply](#).

SHORT COMMUNICATION

Recent recommendations of the Consultative Committee for Length (CCL) regarding strategies for evaluating key comparison data

Jennifer E Decker^{1,4}, N Brown², Maurice G Cox³, A G Steele¹ and R J Douglas²

¹ Institute for National Measurement Standards (INMS), National Research Council of Canada (NRC), Ottawa K1A 0R6, Canada

² National Metrology Institute Australia (NMIA), Australia

³ National Physical Laboratory, Teddington TW11 0LW, UK

E-mail: Jennifer.Decker@nrc-cnrc.gc.ca

Received 6 July 2006

Published 6 November 2006

Online at stacks.iop.org/Met/43/L51

Abstract

A workshop on statistical techniques for evaluating key comparison data in length metrology was held as part of the September 2005 meetings of the Working Group on Dimensional Metrology of the Consultative Committee for Length (CCL) of the International Committee for Weights and Measures. This paper summarizes the discussion at this workshop and the resulting recommendations, subsequently adopted by the CCL, for evaluating future key comparison data in dimensional metrology. Motivation and supporting information relating to the recommendations are included.

1. Introduction

Many key comparisons in dimensional metrology have been completed to date, yielding the opportunity to profit both from the experience of piloting and evaluating key comparison data, and from recent communications concentrating specifically on statistical methods for evaluating *Comité international des poids et mesures* (CIPM) key comparison data [1–5]. One of the goals of the Working Group on Dimensional Metrology (WGDM) of the CIPM Consultative Committee for Length (CCL) is to agree on a defensible general-purpose methodology that can be used to evaluate key comparison data without requiring the Group to make recommendations regarding the procedure to be used in each individual case. To achieve this end, an international workshop on statistical techniques for evaluating key comparison data in length metrology was held as part of the September 2005 meetings of the CCL-WGDM. This paper outlines the steps

for dimensional metrology key comparison data evaluation recommended by the CCL-WGDM and supported by the CCL, including a summary of the workshop discussion together with the motivation and supporting information relating to the recommendation.

2. Guidelines

All key comparisons performed in support of the CIPM Mutual Recognition Arrangement (MRA) [6] follow established general guidelines [7] for selecting, conducting and evaluating key comparisons. CCL-WGDM workshop participants agreed that detailed approaches for comparison data analysis based on generally accepted and broadly used statistical methods were preferable to the introduction of case-specific ad hoc or post hoc schemes. The main reason for this preference is the wealth of theoretical underpinning that exists for these methods, which is not generally available for the other

⁴ Author to whom any correspondence should be addressed.

schemes. To this end, the following guidelines were developed by consensus.

- (i) Determine the inverse-variance weighted mean based on the measured values and associated standard uncertainties submitted by the participants. If no participant provided a (finite) degrees of freedom associated with its declared uncertainty, perform a conventional chi-squared consistency test for this weighted mean and the data. Otherwise, perform an extended chi-squared test [8].
- (ii) (a) If the consistency test is satisfied at the 5% level, use the weighted mean as the key comparison reference value (KCRV) and the standard uncertainty associated with it as the standard uncertainty associated with the KCRV.
 - (b) If the consistency check fails at the 5% level do the following.
 - (1) Determine the largest subset of participants' results that are consistent (at the 5% level) according to the conventional or extended chi-squared test, as appropriate.
 - (2) The pilot alerts participants whose results are not contained in the largest consistent subset that there may be problems with their data. The participants try to determine technical reasons for the laboratories' inconsistent results (blunders, inappropriate corrections for systematic effects, method differences, etc).
 - (3) At the discretion of the participants, the pilot could perform additional modelling (e.g. accounting for drift or travelling artefact instability) in order to increase the size of the largest consistent subset;
 - (4) Use the weighted mean of the largest consistent subset as the KCRV and the standard uncertainty associated with it as the standard uncertainty associated with the KCRV.
- (iii) Derive unilateral and bilateral degrees of equivalence for all participants, and publish the results of the comparison, providing appropriate advice about using the determined KCRV and the associated standard uncertainty.

3. Discussion

Several statistical approaches for obtaining a KCRV determined as some linear combination of the measured values provided by the participants were considered. These approaches concentrated on the determination of the weighting factors in the linear combination from the spread of participants' results [9–11]. Such approaches, which can sometimes offer insight into problems experienced during the comparison, are essentially ad hoc in character, and hence difficult to justify in a formal manner. Any post hoc adjustment of weighting factors made when forming the KCRV can make it hard to establish a clear link between the participants' declared uncertainties, as reflected in their uncertainty budgets, and the uncertainty associated with the KCRV.

Moreover, these approaches are not *model-based*, i.e. they do not directly relate, in a functional manner, input quantities (of which the participants' measured values are realizations)

to an output quantity (estimated by the KCRV). Consequently, they possess an inherent difficulty: the principles of the *Guide to the Expression of Uncertainty in Measurement* (GUM) [12] cannot straightforwardly be applied to evaluate the uncertainty associated with the KCRV, and hence unilateral degrees of equivalence (DoEs) cannot be determined easily.

Some CCL key comparisons give rise to particular data evaluation challenges. In the CCL-K2 comparison of the central length of long gauge blocks [13], uncertainty components associated with changes in the artefact property that occurred during the comparison were included in the uncertainty budgets for purposes of evaluation of DoEs. For some other CCL key comparisons, a number of measured values or coordinates are provided by each participant. One example is the coordinate measurements of 25 balls constituting a two-dimensional array on a ball plate, and another is the line scale comparisons. The measured results may need to be modelled with fixed and length-dependent components [14].

The relative merits of the use of the arithmetic mean and the inverse-variance weighted mean as the KCRV were discussed in the context of these comparisons. (For the inverse-variance weighted mean, the weights are taken to be equal to a constant multiple of the squared reciprocals of the declared standard uncertainties associated with the measured values, the constant being determined such that the weights sum to unity.) The arithmetic mean is sometimes used as a 'politically correct' solution, in that all measured values are treated identically. The weighted mean may be deemed to be justified scientifically in cases where it is consistent with the measurement data (sections 4 and 5).

4. Consistency of results

Of primary interest is whether the observed dispersion in the measured values reported by the participants is consistent with the scatter of values that might be expected based on the associated uncertainty claims. The extent to which the participants' data are consistent with the weighted mean must take account of both the submitted measured values *and* the associated measurement uncertainties declared by the participants. Issues surrounding the assessment of consistency constituted a major consideration during the workshop. Testing the consistency of comparison results based on an appropriate null hypothesis can be carried out and reported conveniently by using chi-squared statistics. Two approaches were discussed in detail during the workshop.

In the conventional approach summarized by Cox [1], the inverse-variance weighted mean, the associated standard uncertainty and the observed chi-squared value, χ_{obs}^2 , are calculated from the participants' results. In a key comparison with N participants, the probability of exceeding χ_{obs}^2 by chance is determined from the distribution function for the chi-squared distribution with $N - 1$ degrees of freedom. Satisfaction of this chi-squared test can be regarded as a statement that the conditions hold for applicability of the weighted mean [1].

When this probability falls below 5%, the hypothesis that, as a whole, the participants' submitted results are consistent with the weighted mean is rejected. In such a case, it is

inappropriate to use the overall weighted mean as the KCRV, and therefore it is not used to form DoEs.

Some consideration was given to more advanced methods for checking consistency, whilst accommodating specifics of the key comparison. In cases where multiple artefacts are circulated as a group, as is routine for gauge block comparisons, systematic effects common to individual NMIs' results can be expected when the same measurement procedure is used for the ensemble of gauge blocks. As a consequence, if evaluated as a group, the results correspond to correlated quantities and hence have an associated covariance matrix (uncertainty matrix) that includes appropriate off-diagonal terms [15–17]. The elements in this covariance matrix can be quantified from the information provided in carefully prepared uncertainty budgets. Solving the generalized least-squares problem, for the parameters of an appropriate measurement equation, incorporates this covariance matrix and is an appropriate way to address this type of covariance when Gaussian distributions are used to underpin the uncertainties of the participants' measured values.

5. Extended tests for consistency

The chi-squared test is based on the assumption that the quantities of which the NMIs' measured values are realizations can be characterized by Gaussian distributions. When the uncertainties associated with one or more of these measured values have related finite degrees of freedom, this assumption can be unreasonable if small numbers of repeated indications have been used for obtaining the expectation and standard deviation. In such situations, and for dealing with non-Gaussian distributions in general, the Monte Carlo approach [18–20] can be applied to extend [5, 8, 21–23] traditional chi-squared testing for more rigorous assessment of consistency. When, as above, there are finite degrees of freedom associated with the declared uncertainties, the distributions that often apply are t -distributions (provided that the repeated indications can be regarded as independently drawn from a Gaussian distribution of unknown standard deviation). These t -distributions have broader tails than Gaussian distributions, and should be used instead in the data evaluation.

In such circumstances, an easy-to-use and readily-available Excel toolkit [24] is applicable to evaluate consistency using Monte Carlo calculations. This toolkit was demonstrated at the workshop, using the results of a recent SIM gauge block comparison [25]. It was shown that, by correctly treating the shape of the probability distributions when computing the chi-squared-like statistic appropriate to a specific comparison, the probability of exceeding χ_{obs}^2 is greater than that obtained when infinite degrees of freedom are assumed. This approach offers the advantage that demonstration of consistency is more closely related to the actual claims of the key comparison participants. In particular, inconsistency is less likely if proper account, as above, is taken of any finite degrees of freedom related to declared participant uncertainties. Related worked examples based on CCL key comparison data previously discussed at the WGDM are currently in preparation for publication.

In addition to this extended chi-squared test of consistency with the weighted mean, the demonstrated toolkit evaluates

extended chi-squared consistency [5] with two other KCRV candidates (the arithmetic mean and the median). It also reports extended chi-squared tests of an aggregated measure of bilateral consistency without invoking any particular KCRV [21]. Nevertheless, there was a general preference for the inverse-variance weighted mean as the KCRV when all participants' results are consistent with it.

6. Treating inconsistency

When the key comparison results fail conventional, extended or any other consistency test that had been agreed upon by the participants, it is not appropriate to use the weighted mean of all participants as a KCRV. In this circumstance, DoEs calculated relative to such a KCRV are not considered to be meaningful for describing the performance of the participants. The participants could decide to investigate individual results in an attempt to identify discrepant data as erroneous. In a few key comparisons, no KCRV has been deemed necessary or suitable for reporting the performance of the participants. Alternatively, a new measurement model could be introduced in an attempt to account for the inconsistency. Models based on plausible physical causes, such as artefact drift or failure within a set of circulating standards, should be considered first. Examination of the model and its modification towards a more realistic interpretation were generally considered to be preferable to the exclusion of some participants' results when attempting to obtain a consistent data set. If, in a particular key comparison, exclusion of some participants' results from those used to determine an (unacceptable) KCRV is deemed necessary, it is appropriate to consider first if there are physical reasons for excluding particular participants' data (admitted blunders or misinterpretations). Alternatively, consideration can be given to investigating whether some participants' values are being assigned undue weight in the light of either their stated degrees of freedom or the accuracy of the gradient of their claimed or implied probability distribution in the vicinity of the KCRV [26]. Software to determine the largest consistent subset was demonstrated at the workshop. A publication detailing the fundamental concepts and methods for its determination is in preparation [27], and a summary of the approach is included here as an appendix.

Bilateral DoEs can often be computed directly from the participants' submitted data. Cases when such DoEs cannot be formed directly from the bilateral data include comparisons when different artefacts are measured by different subsets of the participants [15–17] or where the artefact exhibited a significant drift or other variation. The use of a bilateral DoE, normalized by the uncertainty associated with that difference (' E_n value'), is one measure of the relative performance of a pair of key comparison participants. When it applies, this approach has the advantage that it does not require the computation of a reference value to serve as a mediating quantity when determining consistency. It therefore avoids any difficult issues with the uncertainty associated with the KCRV and correlations associated with participants' measured values and the KCRV [28]. The Excel toolkit [24] can be used to avoid or handle this covariance, even for non-Gaussian distributions. Using its pair-difference chi-squared-like results, the toolkit can be used iteratively for determining

a consistent subset, since the toolkit does not automatically evaluate nor correct for the iterative increase in the false rejection rate as discussed in Beissner [3]. The demonstrated toolkit generates a rigorous 'normalized deviation' for each participant, aggregated relative to all other participants in the comparison. Compatible toolkits [29] can be used to prepare tables of DoEs.

There is concern in some quarters regarding the possibility that the consistency test is satisfied even when a participant reports an exceptionally small uncertainty, namely, when it is smaller than the accepted state-of-the-art minimum value for such measurement. In a statistical sense, there would be no evidence to doubt that the weighted mean cannot be accepted as a KCRV for the comparison. In a metrological sense, it might be deemed unreasonable by the bulk of the participants that the weighted mean be accepted in this circumstance. Accordingly, the pilot of the key comparison would wish to initiate a dialogue with the participant concerned to try to resolve the difficulty. (Such a small uncertainty might be regarded as unacceptable prior to the data evaluation according to the protocol of the comparison.) An appropriate attitude to take is that a statistical test should supplement, and not replace, an examination of the data based on detailed knowledge of the metrological area concerned.

7. Conclusions

Outcomes of the International Workshop on Statistical Techniques for Evaluating Key Comparison Data in Length Metrology held as part of the September 2005 meetings of CCL-WGDM are described. The Group agreed upon a defensible general-purpose methodology to evaluate key comparison data without requiring the CCL-WGDM to make recommendations regarding the procedure for each individual case. The CCL-WGDM then recommended steps for evaluation of key comparison data in dimensional metrology, which were subsequently supported by the CCL. The steps are listed, followed by a general discussion of statistical consistency of comparison results as it relates to the recommended guidelines.

Acknowledgment

The authors appreciatively acknowledge the interest and support of Andrew Wallard, Director of the BIPM. The National Measurement System Directorate of the UK's Department of Trade and Industry supported the work of the National Physical Laboratory in this area.

Appendix A. The largest consistent subset

Consider the simple circulation of a single stable travelling standard around the NMIs participating in a key comparison. Consider the set of data consisting of a measurement result, comprising a measured value and the associated standard uncertainty, provided independently by each such NMI. Each measurement result is regarded as the corresponding NMI's best estimate of the stipulated property of the artefact. The inverse-variance weighted mean of these measurement results is formed. If this weighted mean is consistent with the

measurement results according to a chi-squared measure, it can be accepted as a KCRV for the comparison. Otherwise, the largest consistent subset of the measurement results, i.e. a subset from the complete set that corresponds to as many participating NMIs as possible and that is consistent with the weighted mean of the subset, is determined. This subset is not in general equal to that which would be obtained by successively excluding the most discrepant measurement result. Reference [27] describes an efficient approach, based on the properties of the chi-squared function, for determining the largest consistent subset having the smallest chi-squared value, and applies it to measurement results from several key comparisons.

References

- [1] Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589–95
- [2] Beissner K 2002 On a measure of consistency in comparison measurements *Metrologia* **39** 59–63
- [3] Beissner K 2003 On a measure of consistency in comparison measurements: II. Using effective degrees of freedom *Metrologia* **40** 31–5
- [4] Kacker R, Datla R and Parr A 2002 Combined result and associated uncertainty from interlaboratory evaluations based on the ISO Guide *Metrologia* **39** 279–93
- [5] Steele A G and Douglas R J 2005 Chi-squared statistics for KCRV candidates *Metrologia* **42** 253–61
- [6] <http://www.bipm.org/en/cipm-mra/>
- [7] <http://www.bipm.org/utis/en/pdf/guidelines.pdf>
- [8] Steele A G and Douglas R J 2006 Extending chi-squared statistics for key comparisons in metrology *J. Comput. Appl. Math.* **192** 51–8
- [9] Thalmann R 2002 CCL key comparison: calibration of gauge blocks by interferometry *Metrologia* **39** 165–77
- [10] Stone J A 2005 Methods for evaluating the reference value in laboratory intercomparisons of dimensional measurements *Recent Developments in Traceable Dimensional Measurements III* ed J E Decker and G S Peng *Proc. SPIE* **5879** 58790V–1–8
- [11] Brown N 2005 A new approach to determining the key comparison reference value *Recent Developments in Traceable Dimensional Measurements III* J E Decker and G S Peng ed *Proc. SPIE* **5879** 58790W-1-7
- [12] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP and OIML 1995 *Guide to the Expression of Uncertainty in Measurement* (Geneva: International Organization for Standardization) corrected and reprinted
- [13] Lewis A 2003 CCL-K2: Long gauge block measurement by interferometry: final report *Metrologia* **40** Tech. Suppl. 04004
- [14] Bosse H, Häbeler-Grohne W, Flügge J and Köning R 2003 Final report on CCL-S3 supplementary line scale comparison Nano3 *Metrologia* **40** Tech. Suppl. 04002
- [15] Cox M G, Harris P M and Woolliams E R 2005 Data evaluation of key comparisons involving linked bilateral measurements and multiple artefacts *NCSL Int. Workshop Symp. (Washington, USA)*
- [16] Cox M G, Harris P M and Woolliams E R 2006 Data evaluation of key comparisons involving several artefacts *Advanced Mathematical Tools in Metrology VII* ed P Ciarlini *et al* (Singapore: World Scientific) pp 23–34
- [17] Woolliams E R, Fox N P, Cox M G, Harris P M and Harrison N J 2006 The CCPR K1—a key comparison of spectral irradiance from 250 nm to 2500 nm: measurements, analysis and results *Metrologia* **43** S98–104
- [18] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML 2006 Evaluation of measurement data—*Supplement 1 to the Guide to the Expression of Uncertainty in*

- Measurement— Propagation of Distributions Using a Monte Carlo Method* (Joint Committee for Guides in Metrology) Draft
- [19] Steele A G and Douglas R J 2006 Simplifications from simulations *NCSLI Meas.* **1** 57–68
- [20] Steele A G and Douglas R J 2005 <http://inms-ienm.nrc-cnrc.gc.ca/qde/montecarlo/>
- [21] Douglas R J and Steele A G 2006 Pair-difference chi-squared statistics for key comparisons *Metrologia* **43** 89–97
- [22] Steele A G and Douglas R J 2006 Simplicity with *advanced mathematical tools* for measurement and testing *Measurement* **39** 795–807
- [23] Steele A G and Douglas R J 2006 Extending E_n for measurement science *Metrologia* **43** S235–43
- [24] Steele A G and Douglas R J <http://inms-ienm.nrc-cnrc.gc.ca/qde/montecarlo/EnToolkit.html>
- [25] Decker J E, Alschuler J, Candanedo C J, De la Cruz E, Esteban E P, Morales R, de Oliveira J C V, Stone J, Stoup J and Pekelsky J R SIM.4.2 Regional Comparison Stage One: Calibration of gauge blocks by optical interferometry *Metrologia* submitted
- [26] Steele A G, Wood, B M and Douglas R J 2005 Outlier rejection for the weighted-mean KCRV *Metrologia* **42** 32–8
- [27] Cox M G The evaluation of key comparison data: determining the largest consistent subset. In preparation
- [28] Steele A G, Wood B M and Douglas R J 2001 Exclusive statistics: simple treatment of the unavoidable correlations from key comparison reference values *Metrologia* **38** 483–8
- [29] Wood B M, Steele A G and Douglas R J 2000–2003 <http://inms-ienm.nrc-cnrc.gc.ca/qde>