XVIII IMEKO WORLD CONGRESS
Metrology for a Sustainable Development
September, 17 – 22, 2006, Rio de Janeiro, Brazil

# EVALUATING RESULTS OF INTERNATIONAL COMPARISONS: WORKED EXAMPLE OF CCL-K2 COMPARISON OF LONG GAUGE BLOCK CALIBRATION

*J .E. Decker* [1], *A. J.Lewis*[*2], *M .G. Cox** [3], *A. G. Steele* [4], *R. J. Douglas* [5]

Institute for National Measurement Standards, National Research Council, Ottawa, Canada
[1] jennifer.decker@nrc-cnrc.gc.ca
[4] alan.steele@nrc-cnrc.gc.ca
[5] rob.douglas@nrc-cnrc.gc.ca
*National Physical Laboratory, Teddington, UK
[2]Andrew.Lewis@npl.co.uk
[3] Maurice.Cox@npl.co.uk

**Abstract:** A method for the evaluation of international comparison results is demonstrated via detailed consideration of the CCL-K2 Key Comparison of long gauge block calibration. Although CCL-K2 involves a geometrically simple length measurand, associated with it are many of the difficult measurement issues often encountered in international comparison exercises.

**Keywords:** key comparison, uncertainty evaluation, gauge block

## 1. INTRODUCTION

Key comparisons test the compatibility of principal measurement capabilities in all major metrology areas. The evaluation of key comparison data poses special problems when conventional testing indicates that the inverse variance weighted mean is inconsistent with the participants' measurement data.

At the September 2005 meeting of the Consultative Committee for Length (CCL), a general procedure for the analysis of key comparison results was adopted [1], which includes the use of a 'toolkit' for evaluating equivalence. We demonstrate the En and QDE Toolkits, and highlight the results of their application to a well-known CCL key comparison in long gauge block calibration of central length by the method of optical interferometry (CCL-K2) [2] using a variety of chi-squared statistical techniques. Special attention is paid to the uncertainty claims of the participants, with emphasis on the implications of quoting finite degrees of freedom to the chi-squared-like distribution appropriate for making decisions on consistency [3, 4]. This paper provides a practical demonstration of the En and QDE Toolkits, whereas pointed technical discussions of comparison results are deferred to the published report. The evaluation below follows the guidelines developed at the September 2005 meeting of the CCL-WGDM [1].

## 2. TOOLKIT DEMONSTRATION

Both the En Toolkit and the QDE Toolkit for evaluation of measurement comparison results can be downloaded from the website *http://inms-ienm.nrc-cnrc.gc.ca/qde/*, where instructions for their application appear.

### 2.1. Evaluation of Key Comparison Data

The evaluation starts with the data constituting, for each participant, a measurement value, and the associated standard uncertainty and degrees of freedom provided in a Microsoft Excel spreadsheet. See Table 1 for the example of the 175 mm (S/N 6071) gauge block.

The En Toolkit consists of Excel macros with which to evaluate the inverse-variance weighted mean, simple arithmetic mean, and median, their associated uncertainties and degrees of freedom. In the Excel macro pop-up list, each Toolkit macro is accompanied by a brief description, and the Toolkit automatically distributes comments that identify the contents of cells resulting from the evaluations performed by the macros. Table 2 lists results for values of central tendency (as candidate reference values) after running the *En_TableBuilder* macro[1] on the data set of Table 1. The Toolkits employ Monte Carlo methods. The uncertainty associated with the weighted mean is evaluated using the recognized formula [5], extended to cover the cases where a covariance matrix has been entered. That associated with the unweighted mean is evaluated this way after setting the weights top be identical. Effective degrees of freedom are obtained using the Welch-Satterthwaite approximation. The median is instance median, and a comment at this cell also gives the mean of the re-sampled distribution of medians [5] and the associated standard uncertainty is the standard deviation of that distribution.

[1] The average run time of this macro is about 10 minutes on a Pentium computer with a 1.5 GHz processor.

Table 1. Values of gauge block deviation from nominal central length, standard uncertainty and degrees of freedom as reported by participants entered in the format for running En Toolkit macros.

| Lab | $\Delta L$ /µm | $u_c$ /µm | $\nu_S$ |
|---|---|---|---|
| IMGC | 0.140 | 0.028 | *65* |
| PTB | 0.122 | 0.013 | *85* |
| NPL | 0.161 | 0.030 | *241* |
| NIST | 0.142 | 0.016 | *50* |
| INMETRO | 0.150 | 0.020 | *100* |
| NRC | 0.125 | 0.027 | *150* |
| NRLM | 0.148 | 0.019 | *10* |
| NIM | 0.194 | 0.019 | *40* |
| CSIRO | 0.154 | 0.023 | *195* |
| CSIR | 0.180 | 0.110 | *39* |
| SMU | 0.285 | 0.038 | *210* |
| VNIIM | 0.312 | 0.021 | *8* |

Table 2. Results of En Toolkit evaluation of reference values (RV) for variance weighted mean (W-Mean), simple arithmetic mean (S-Mean), and median.

|  | Value /µm | Standard Uncertainty /µm | Degrees of Freedom |
|---|---|---|---|
| *RV S-Mean* | 0.176 | 0.011 | 89.3 |
| *RV W-Mean* | 0.163 | 0.006 | 286.8 |
| *RV Median* | 0.152 | 0.011 | — |

The classical chi-squared test is used to address the question of metrological equivalence, namely that the laboratory measurement results have the same mean, and that the dispersion of results is adequately described by the measurement uncertainties stated by the participants. This is the conventional null hypothesis, supplemented with the assumption that the inverse-variance weighted mean adequately describes the common mean. The null hypothesis consistency test suggested in the literature [5] for application to key comparison data evaluation calculates the probability that the observed value of chi-squared exceeds a critical value of chi-squared by chance. Specifically, if $\Pr\{\chi^2(\nu) > \chi_{obs}^2\} < 5$ %, the consistency check fails, and some alternative or supplementary description will be required. Otherwise, publication of the comparison can proceed without any additional description.

The *En_TableBuilder* macro automatically provides values for $\chi_{obs}^2$ and $\Pr\{\chi^2(\nu) > \chi_{obs}^2\}$ based on participant data such as that listed in Table 1. Using this example, including all participants, and following conventional practice [5] based on the weighted mean [2], the data set is not consistent. These values of $\chi_{obs}^2$ and $\Pr\{\chi^2(\nu) > \chi_{obs}^2\}$ are evaluated by running the *En_TableBuilder* macro with the degrees of freedom cells all set to “normal”.

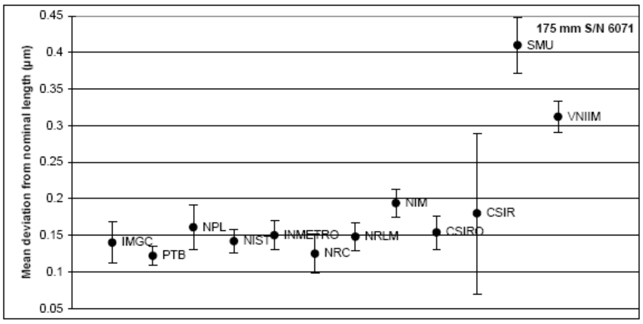

Figure 1. The reported values (Table 1) with bars representing ±1 standard uncertainty for the 175 mm gauge block of the CCL-K2 comparison.

According to recommended CCL guidelines [1], the next step in the procedure for the comparison pilot is to determine the largest subset of participants' results that is consistent. The data for our example are shown plotted in Figure 1, whereupon visual inspection it is clear that there are two measurement results that are far from the cluster of the other results. The largest consistent subset can be determined by an appropriate algorithm [6].

This largest consistent subset does not contain the SMU and VNIIM results. Excluding these results from the data set, and running the conventional consistency test (assuming all the distributions are normal when running the *En_TableBuilder* macro) yields $\Pr\{\chi^2(\nu) > 1.209\} =$ 28.42 %. The consistency check can also be carried out with the En Toolkit taking into consideration the degrees of freedom submitted by the participants. Running the *En_TableBuilder* macro again, this time including the reported values for degrees of freedom, $\Pr\{\chi^2(\nu) > 1.209\} =$ 31.82 % for the same subset of participants. Values are summarized in Table 3. The consistency test passes at the 5 % level, and therefore this subset of participants can be designated a consistent subset. The extended version of the chi-squared testing is observed to be advantageous for the comparison participants in demonstrating consistency because the probability is greater, and both the GUM-compliant standard uncertainty and degrees of freedom that were carefully evaluated by participants are utilized constructively.

At this point in the procedure [1], the pilot alerts participants whose results are not contained in the largest consistent subset that there may be problems with their data, and the participants try to determine technical reasons for the inconsistent results, as discussed in the CCL-K2 report.

[2] The En Toolkit table also includes chi-squared testing against two other common reference value candidates: the simple mean and the median.

Table 3. The 175 mm gauge block data evaluated using the largest consistent subset (the results of two laboratories, SMU and VNIIM were excluded). Results of the consistency check for variance weighted mean applying the conventional evaluation (normal distributions) compared with the extended evaluation, which takes into consideration the degrees of freedom submitted by participants.

| | Conventional testing (normal distributions) | With Participants' submitted degrees of freedom |
|---|---|---|
| $\chi_{obs}^2$ | 1.209 | 1.209 |
| $\Pr\{\chi^2(\nu) > \chi_{obs}^2\}$ | 28.42 % | 31.82 % |
| Weighted mean | 0.1454 μm | 0.1454 μm |
| Associated standard uncertainty | 0.0065 μm | 0.0065 μm |

The QDE Toolkit macro[3] *tk_mraCCQM_TableBuilder* run with input data of the desired participant data set, and the reference value determined from the largest consistent subset creates a table of equivalence in a convenient format for reporting and submission to the BIPM database. One of the advantages of using the Toolkit macros is the simple format of the input data and resulting evaluations for easy checking and proof-reading.

## 3. UNCERTAINTY ASSOCIATED WITH A TRAVELLING ARTEFACT

When results are not statistically consistent, either the largest consistent subset of the participants' data may be determined, or technical reasons sought for the inconsistent results. In CCL-K2, participants evaluated an uncertainty component associated with the wear and change in condition of the material gauge block standard attributed to the time duration of the comparison. The artefact uncertainty component was applied when comparing participants' results with the reference value. This value was not used in the context of statistical consistency testing.

## 4. CONCLUSIONS

Application of macros from the En Toolkit in Microsoft Excel for the evaluation of KC data in accordance the guidelines recommended by the CCL are illustrated via the specific example of the CCL-K2 long gauge block comparison of central length calibration. The statistically robust evaluation of the consistency of a set of measurement values and associated uncertainties by both conventional and extended chi squared testing is introduced and explained. Application of the QDE Toolkit for constructing tables of equivalence used in key comparison reporting is demonstrated.

## REFERENCES

[1] J. E. Decker, N. Brown, M. G. Cox, A. G. Steele, R. J. Douglas, "Recent decisions of the Consultative Committee for Length (CCL) regarding strategies for evaluating key comparison data," *Metrologia*, submitted.

[2] A. Lewis, "Long gauge block measurement by interferometry: Final Report," *Metrologia*, **40**, 04004, 2003.

[3] A. G. Steele, R. J. Douglas, "Chi-squared statistics for KCRV candidates," *Metrologia*, **42**, n°4, 253-261, 2005.

[4] A. G. Steele, R. J. Douglas, "Extending chi-squared statistics for key comparisons in metrology," *Journal of Computational and Applied Mathematics*, **192**, 51-58, 2006.

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

[6] M. G. Cox, The evaluation of key comparison data: determining the largest consistent subset. In preparation.

[3] Builds tables of equivalence with negligible delay.