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

A SURVEY OF KEY COMPARISONS

A Survey of Design, Analysis, and Reporting of Results in Key Comparisons

FULL PAPER FOR THE XVIII IMEKO WORLD CONGRESS

Adriana Hornikova *and* *William F. Guthrie*

NIST, 100 Bureau Dr., Gaithersburg, MD, adriana.hornikova@nist.gov; will.guthrie@nist.gov

Abstract - Key comparisons are international inter-laboratory studies used to establish the degree of equivalence between national measurement standards. These studies, carried out by National Metrology Institutes (NMIs), are time-consuming, but necessary to facilitate international trade. From the signing of the Mutual Recognition Arrangement (MRA) in 1999 through the end of 2004, 85 key comparisons in a wide range of metrological areas were completed and have results posted in the Key Comparison Database (KCDB) maintained by the BIPM in France and in the International Comparisons Database (ICDB) maintained by NIST in the U.S [1,2].

Supported by this large set of completed comparisons from the KCDB and the ICDB, an opportunity has arisen to study the methods that are being used to conduct key comparisons. This paper summarizes work on currently completed key comparisons and offers recommendations for the design, analysis, and interpretation of future comparisons.

Keywords: Inter-laboratory studies, Experiment design, Key Comparison**.**

1. INTRODUCTION

In this paper we discuss a survey of the results of recently completed Key Comparisons, as shown in Table 1, to learn how Key Comparisons are actually being carried out in different metrological areas. The survey included questions on the general organization of the Comparisons and the equipment being used, as well as covering a number of different statistical aspects of the Comparisons, including: Comparison design, types of data collected and reported, method used to calculate a Key Comparison Reference Value, calculation of degrees of equivalence between pairs of participating NMIs, and linkage to different comparisons. Altogether, answers to 28 questions about each of the 85 comparisons were recorded in the survey. The questions from the survey, organized in to general categories are listed below.

General questions:

- Organizing committee (consultative committee)
- Comparison name
- Type of comparison (key, supplementary, EUROMET)
- Years executed
- Pilot laboratory
- Number of participating laboratories
- Measurand

Questions related to transfer artifact:

- Number of transfer artifacts
- Number of transfer artifacts measured per laboratory
- Quality of transfer artifact
- Drift of transfer artifact

Questions about design and Key Comparison Reference Value (KCRV):

- Nominal values
- Number of nominal values
- Type of KCRV, is there a physical basis for a KCRV
- Type of design (balanced or unbalanced)
- Type of experiment (traveling of the transfer artifact)
- Data reported
- Distribution of data and its shape
- Outliers and outlying laboratories, possible disclosure
- Different methods
- Number of replicated measurements and degrees of freedom
- Functional relationship, physical basis of the measurements

Uncertainty related questions:

- Source of type A uncertainty
- Ratio of maximum to minimum reported laboratory uncertainty
- Percentage of laboratories with type B uncertainty dominating
- Possibility of establishing a link and existing linkage
- Other unusual features of the comparison.

Table 1: Summary of Key Comparisons surveyed.

Category	*Number of KCs Surveyed*
Acoustics, Ultrasound and Vibration	4
Amount of Substance	31
Electricity and Magnetism	18
Ionizing radiation	3
Length	6
Mass (and related quantities)	13
Photometry / Radiometry	7
Thermometry	3

The next four sections of the paper highlight some of our findings from the survey on Comparison design, the data presented in final report, comparisons of uncertainties across laboratories, and methods used to compute Key Comparison Reference Values (KCRVs).

2. SELECTED SURVEY FINDINGS

2.1 Comparison Design

One of the initial findings from review of the reported comparison results relates to the descriptions of the designs used to carry out these experiments. The experiment design for a key comparison typically includes the specification of a high-level design that describes the basic type of measurement to be done, a set of nominal measurement conditions, and the order in which one or more transfer artifacts will be circulated among the participating NMIs. Special instructions for the use of the transfer artifact may also be given to avoid damage to the transfer artifact or to help ensure comparability of the results. In some cases an optional lower-level design that specifies further details, such as the exact measurement conditions each laboratory will use, the number of measurements each laboratory will make under each set of measurement conditions, or the order in which the measurements should be made, may also be a part of the complete comparison design. In most comparisons, however, only the high-level design and the special instructions for the use of the transfer artifact are set in advance and the low-level design is individually set by each participant.

To illustrate, one of the simplest high-level designs that might be used in a key comparison is one in which one transfer artifact is sent sequentially from the pilot laboratory, which is organizing the study, to each of the other laboratories participating in the study. In other studies several artifacts are used so different laboratories can make measurements in parallel. While not complex as planned, the actual designs used in most comparisons often turn out to be complicated, because transfer artifacts are damaged during shipment or use, alternate artifacts must be incorporated in the design to accommodate special conditions at one or two laboratories, or due to use of multiple measurement standards or multiple transfer artifacts at some laboratories, etc.

Despite these complications, however, most reports do not include easy to understand descriptions of the actual designs used. Most authors provide verbal descriptions of the design or a table listing the dates when each transfer artifact was sent to each NMI. However, these tables do not always provide a clear indication of which artifacts were actually used by each laboratory. Even when they do indicate the usage of specific transfer artifacts, a fair amount of effort is often required to see the global pattern of the comparison from a table. *To make it easier to understand the design, the use of diagrams similar to those shown in Figures 1, 2, and 3 is suggested, in addition to the usual verbal or tabular descriptions of the design.*

Following rules were developed to construct these diagrams of comparison designs:

- Each pilot laboratory or sub-coordinating laboratory is in a white oval with the abbreviation of the name of the NMI. Small rectangles in the oval indicate the name or serial number of each transfer artifact the pilot laboratory measures. The rectangle for each transfer artifact is assigned its own color to make it easy to see which laboratories used common transfer artifacts.
- Each laboratory (NMI) is represented by a large rectangle filled with smaller rectangles that indicate (by color) which transfer artifacts have been measured at that NMI.
- Following the time-line of the comparison, a sequence of arrows connecting the laboratories indicates the transfer artifact's progression.

It is also important to use bright, contrasting colors that can be easily distinguished, like blue, green, red and yellow, to build the main color palette, although extremely vivid colors may not work best.

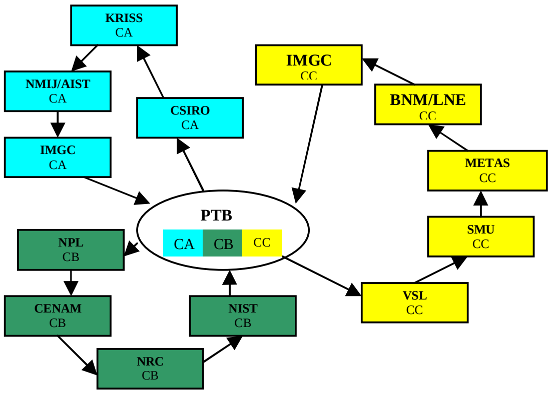


Figure 1: Some comparisons, like this mass comparison (CMM.M-K2), use multiple sets of transfer artifacts with different nominal values. After each sub-group of laboratories in a common loop makes measurements, the results are combined using the results from the pilot laboratory (PTB).

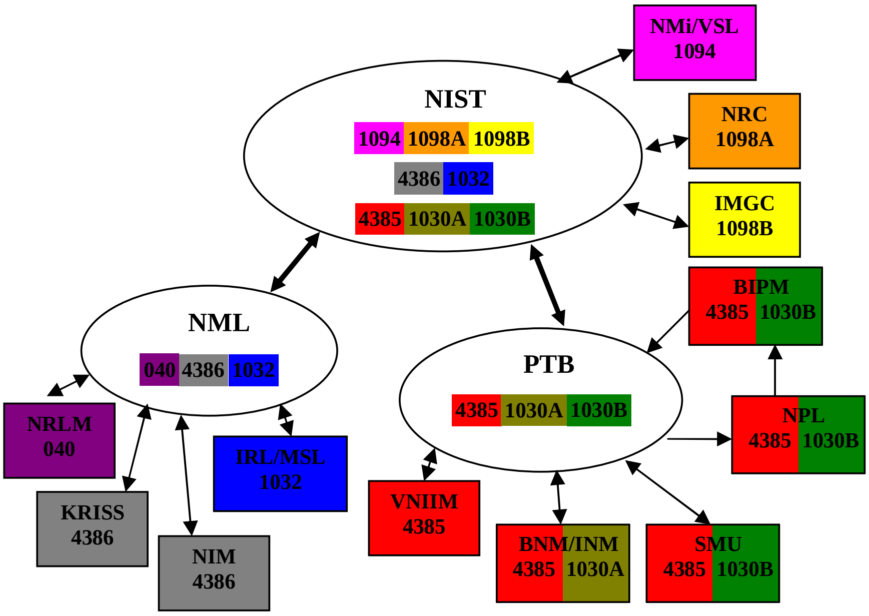
CA: PTB (?) – CSIRO (?) – KRISS (?) – NMIJ/AIST (?) – IMGC (?) – PTB (?)
CB: PTB (?) – NPL (?) – CERAM (?) – NRC (?) – NIST (?) – PTB (?)
CC: PTB (?) – VSL (?) - SMU (?) – METAS (?) – BNM/LNE (?) – IMGC (?) – PTB (?)

Figure 2: Relative time-lines for each transfer artifact used in the comparison shown in Figure 1. This time-line does not show the length of time the transfer artifact was used at each lab or artifact travel times, but those could be easily added if necessary. The numbers in parentheses after each lab would ideally show the number of measurements made.

In addition to schematic drawing in Figure 1 that provide a top-down look at the organization of a comparison, a time-line for the set of transfer artifacts used in a comparison can also be useful. Figure 2 shows the time-lines for each transfer artifact used in the comparison from Figure 1. Each time-line lists the laboratories that measured the artifact in sequence and ideally would also include the number of measurements made by each laboratory in parentheses. The time-lines provide a detailed, but quickly understandable, look at some of the details of the design. Figure 3, below, shows a diagrammatic view and a timeline with the numbers of measurements taken at each laboratory for a more complicated comparison than that shown in Figure 1 and Figure 2.

2.2 Data Reported

The data reported by individual laboratories participating in a key comparison typically includes the mean value obtained from measuring the transfer artifact and an associated estimate of its uncertainty. The uncertainty estimate accounts for the uncertainty in the laboratory's measurements of the transfer artifact as well as the uncertainty in the calibration of the laboratory's measurement equipment using their national measurement standards.



1094: NIST(3) – Nmi-VSL(1)– NIST(1)
1098A: NIST(4) – NRC(1) – NIST(1)
1098B: IMGC(1 or 3) – NIST(2)
4386: NIST(3) – NML(3) – KRISS(1) – NML(1) – NIM(2) – NML(1) – NIST(1)
1032: NIST(3) – NML(1) – MSL(2) – NML(1) – NIST(1)
040: NML(1) – NRLM(4) – NML (1)
4385: NIST(3) – PTB(1) – VNIIM (3) – PTB(1) – BNM(3) – PTB(1) – SMU(3) – PTB(1) – NPL(2) – BIPM(2) – PTB(1) – NIST(1)
1030A: NIST(3) – PTB(1) – BNM(4) – PTB(2)
1030B: SMU(3) – NPL(3) – BIPM(2) – PTB(1) – NIST(2)

Figure 3: This figure illustrates a key comparison in temperature (CCT-K3) with a diagrammatic view and a time line together. In many cases the pilot laboratory needs to re-measure the traveling artifact periodically (e.g. after each laboratory or after several laboratories in a particular region), to check for drift in the transfer artifact. In these situations, multiple loops appear in the design. These loops can be run in parallel if multiple transfer artifacts are used. Partitioning a comparison into regions coordinated by different laboratories, as indicated by the smaller ovals in this diagram, makes the amount of work for each laboratory more manageable.

The individual uncertainty estimates from each source of uncertainty included in the final uncertainty estimate are provided in a table typically called an uncertainty budget. The uncertainty budgets also indicate which individual uncertainty estimates for each source of uncertainty were estimated using statistical methods (denoted type A sources of uncertainty) and which were estimated based on expert opinion (denoted type B sources of uncertainty). Very often, however, the uncertainties are given in terms of a single measurement result (rather than the mean) and the sizes of the samples used for calibration and measuring the transfer artifact are not reported. Similarly, in cases where the uncertainty estimates are not based on data from the comparison, but are obtained from control charts of similar measurements, it is rare for the degrees of freedom associated with those uncertainty estimates to be given.

Without reporting the appropriate sample sizes and degrees of freedom associated with each set of measurements, it is difficult for the data collected in the comparison to be used in any future computations. In particular, this may make the linkage of key comparison results with later regional comparisons (that include more laboratories) difficult or impossible. *As a result, we suggest that the data reported for a key comparison either include all of the raw data, summaries of the data that include the appropriate sample sizes and degrees of freedom, or, ideally, both.*

2.3 Comparison of Uncertainties Across Laboratories

One of the less generally appreciated results from a key comparison is information on how uncertainties compare across laboratories, as shown in Figures 4 and 5.

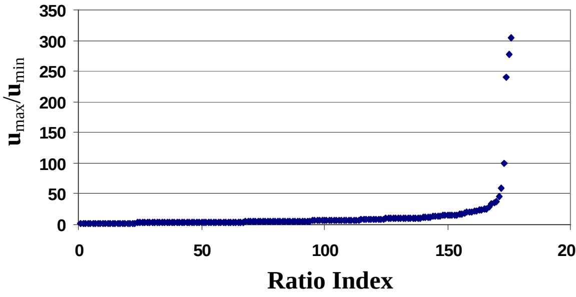


Figure 4: Ratios of the largest to smallest reported combined standard uncertainties, u_C, obtained individually for each nominal value within comparisons. The plot shows ratios plotted in order sorted from lowest to highest.

In many comparisons, the ratio of the largest reported combined standard uncertainty to the smallest is less than 4 or 5, which may be reasonable. There are quite a few comparisons, however, where the ratio is in the range of 10 to 15, and in the most discrepant case, the laboratory with the largest reported uncertainty has a combined standard uncertainty more than 300 times larger than the laboratory from the same comparison with the smallest reported combined standard uncertainty.

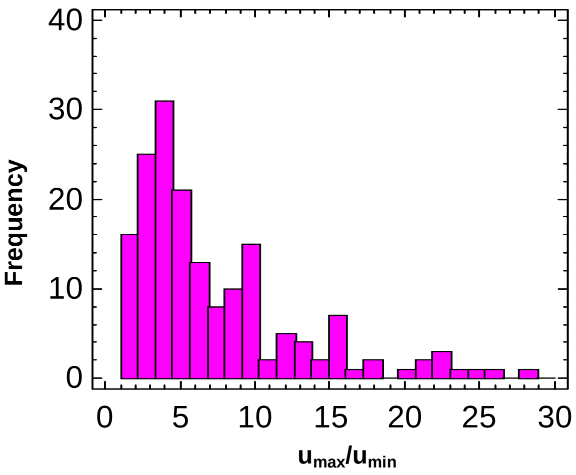


Figure 5: Ratios of the largest to smallest reported combined standard uncertainties obtained individually for each nominal value within each comparison (ratios greater than 30 not shown in this plot to highlight typical values).

This suggests that outliers may be influencing some of the results or that there must be some improved measurement practices that laboratories with relatively large uncertainties could adopt to reduce their uncertainties (based on the experiences of laboratories with smaller uncertainties). Another reason for such a big discrepancy between laboratory uncertainties could be that the uncertainty estimates from different laboratories may have not been estimated on an easily comparable basis. *Further investigation of these issues is needed.*

2.4 Key Comparison Reference Values (KCRVs)

A KCRV is a consensus mean of transfer artifact values measured by participating laboratories. The KCRV is used as a reference to identify laboratories that agree with one another. The use of a KCRV allows a concise comparison between the laboratories since it compares each laboratory to the reference value in order to define a set of homogeneous laboratories. Of course the use of a KCRV to assess equivalence of laboratories does not identify sub-groups of laboratories that are equivalent to each other, but disagree with the majority of the other laboratories, nor does it provide as specific an assessment of the equivalence of any particular pair of laboratories as a pair-wise comparison would do. Questions of interest in the survey related to the estimation of a KCRV include:

- what type of KCRV is used in the comparison (weighted mean, median, none, etc.),
- whether there is physical interpretation for the KCRV, and
- whether or not outlying laboratories were disregarded in the computation of the KCRV.

One of the most popular ways to estimate the KCRV is to use a weighted mean of each laboratory's assessment of the transfer artifact's value, with weights inversely proportional to the uncertainty of the laboratory's estimate of the transfer artifact's value. One potential problem with the use of this estimator, of course, is its reliance on estimated weights that may be based on only a small amount of data. Other estimators that are popular include the equally-weighted mean and the median. Other approaches that have been used to obtain a KCRV in different comparison include the following:

- using the pilot laboratory's value as the KCRV (e.g. using the BIPM value),
- using the nominal value associated with the transfer artifact, and
- using a statistical estimator based on a subset of laboratories maintaining primary reference standards or meeting some other non-statistical criteria.

In cases where comparisons are being made across multiple nominal values, such as different levels of pressure, that follow a physical or empirical statistical model, the KCRV is sometimes defined to be a function that is estimated using a linear fit. Similarly, in cases in which the transfer artifact is subject to drift, a linear least squares regression model with time as the predictor variable is often fit to the pilot laboratory's data and used as the reference function.

One issue that arises from the multiplicity of ways to estimate the KCRV is determining which estimator is most appropriate in a given situation. KCRV estimates obtained using different estimators can lead to qualitatively different results that identify different subsets of laboratories as equivalent. *This problem clearly requires additional research and guidance on the best techniques for use in different scenarios. Some research has been done on this topic [7-15] and more is underway.*

Another more fundamental problem that has arisen with regard to the use of reference values in key comparisons is the physical interpretation of the KCRV. In a number of key comparisons there has been disagreement among the participants over what quantity the KCRV is actually estimating and whether or not a KCRV should be used if it is not an estimate of a well-defined physical quantity. This issue has been resolved in different ways in different comparisons within particular metrological areas. For example, in thermometry there are three different comparisons in which the participants agreed that the KCRV did not have a well-defined physical interpretation, and came to different conclusions about the use of a KCRV in each particular comparison. In comparisons CCT-K2 and CCT-K4 the participants used a weighted mean as the KCRV, but stated that the KCRV had no uncertainty by definition, since it does not represent an estimate of a physically-interpretable quantity. In comparison CCT-K3, however, the participants ultimately agreed not to use a KCRV at all, since it was thought likely the KCRV might be misinterpreted as a redefinition of the International Temperature Scale adopted in 1990. *This is another area in which further research, in collaboration with scientists from each metrological area, is needed so guidelines on when a KCRV should or should not be used can be developed.*

3. CONCLUSIONS

Making systematic use of the international metrological community's current experience with key comparisons through this survey of reported results will help ensure that future key comparisons can be carried out using the best techniques available and will provide an improved basis for mutual recognition of measurement results and international trade.

In particular, to help make the results of key comparisons simpler to understand and as useful as possible, we propose the following guidelines:

- using graphical summaries of the comparison to make the overall design of the comparison outlined in the final report easier to read and more understandable,
- using simple designs with built-in redundancy (in case a transfer artifact breaks),
- reporting all of the data (and other information) necessary to make future computations possible,
- sharing measurement procedures and comparing methods used to estimate uncertainties to ensure that results are comparable and laboratories benefit from their mutual experiences with these measurement procedures,
- working with participants in advance to determine whether or not using a KCRV has a well-defined physical interpretation and makes sense for a given comparison, and
- learning about the statistical assumptions and properties of different KCRV estimators to ensure that the estimator used is appropriate for the type of data at hand.

Other items that have been examined in this survey include the properties of typical comparison designs, the relationship between the Type A and Type B uncertainties in key comparison results, the variation in uncertainty levels among comparison participants, and opportunities for using multivariate data analyses to reduce comparison uncertainties, etc. In addition to continuing to review published reports and results, we also hope to meet with scientists who have participated in different comparisons to learn about other issues that they may have encountered so we can continue to learn how key comparisons are being carried out and how the process may be able to be improved.

REFERENCES

[1] BIPM Key Comparison Database: http://kcdb.bipm.org/AppendixB/KCDB_ApB_search.asp

[2] International Comparisons Database: http://icdb.nist.gov/

[3] Technical Supplement to *Metrologia*: http://www.bipm.org/metrologia/TechSupp.jsp

[4] Guidelines for CIPM key comparisons at http://kcdb.bipm.org/default.asp

[5] NIST Position Statement on the Conduct of Key Comparisons: http://icdb.nist.gov/policies.asp.

[6] NIST Statement on Statistical Principles for the Design and Analysis of Key Comparisons: http://icdb.nist.gov/policies.asp.

[7] Rukhin, A. L., Strawderman, W. E., Statistical Aspects of Linkage Analysis in Inter-laboratory Studies, *Journal of Statistical Planning and Inference*, submitted.

[8] Zhang, N. F., Sedransk, N., and Jarrett, D. G., Statistical Uncertainty Analysis of Key Comparisons CCEM-K2, *IEEE Transaction Instrumentation & Measurement*, 2003, pp. 491-494.

[9] Rukhin, A. L., Two Procedures of Meta-analysis in Clinical Trials and Inter-laboratory Studies, *Tatra Mountains Mathematical Publication*s, 2003, 26, pp. 155-168.

[10] Iyer, H. K., Wang, C.M., and Vecchia D. F., Consistency Tests for Key Comparison Data, *Metrologia*, 2004, vol. 41, pp. 223-230.

[11] Toman, B., Bayesian Approach to Calculating a Reference Value, *Proceedings of NCSL*, 2004.

[12] Wang, C. M., Iyer, H., Generalized Inference for Uncertainty Evaluation with Application to Calibration Experiments, *2004 Proceedings of ASPE Summer Topic Meeting*, 2004.

[13] Zhang, N. F., Liu, H. K., Sedransk, N., Strawderman, W. E., Statistical Analysis of Key Comparisons with Linear Trends, *Metrologia*, 2004, vol. 41, pp. 231-237.

[14] Zhang, N. F., Liu, H. K., Sedransk, N., Strawderman, W. E., Uncertainty Analysis of Inter-laboratory Studies with Linear Trends, *2004 Proceedings of ASPE Summer Topic Meeting - Uncertainty Analysis in Measurement and Design*, 2004, pp. 100-105.

[15] Iyer, H.K., Wang, C.M., Mathew T., Models and Confidence Intervals for True Values in Inter-laboratory Trials, *Journal of the American Statistical Association*, submitted.

Links to final reports of key comparisons:

CMM.M-K2:
http://kcdb.bipm.org/appendixB/appbresults/ccm.m-k2/ccm.m-k2.pdf
http://kcdb.bipm.org/appendixB/appbresults/ccm.m-k2/ccm.m-k2_final_report.pdf

CCT-K2:
http://kcdb.bipm.org/AppendixB/appbresults/cct-k2/cct-k2_final_report.pdf

CCT – K3:
http://kcdb.bipm.org/AppendixB/appbresults/cct-k3/cct-k3.pdf

CCT-K4:
http://icdb.nist.gov/results_descript/AppBResults/pdf/CCT-K4_icdb.pdf
http://kcdb.bipm.org/AppendixB/appbresults/cct-k4/cct-k4_final_report.pdf