MODEL BASED UNCERTAINTY ANALYSIS IN INTER-LABORATORY COMPARISONS

BLAZA TOMAN

Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA

ANTONIO POSSOLO

Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA

Statistical analysis of inter-laboratory comparisons (e.g. Key Comparisons, Supplemental Comparisons) is required to produce an estimate of the measurand called a reference value and further, measures of equivalence of the participating laboratories. Methods of estimation of the reference value have been proposed that rest on the idea of finding a so-called consistent subset of laboratories, that is, eliminating outlying participants. In this paper we propose an alternative statistical model, one that accommodates all of the participants' data and incorporates the dispersion among the laboratories into the total uncertainty of the various estimates. This model recognizes the fact that the dispersion of values between laboratories often is substantially larger than the measurement uncertainties provided by the participating laboratories. We illustrate the method on data from key comparison CCL-K1.

1. Introduction

Data from inter-laboratory comparisons (e.g. Key Comparisons, Supplemental Comparisons) is collected to produce information about a particular measurand, or a set of measurands. In the case of key comparisons, the primary goal is to produce measures of equivalence of the participating laboratories, both unilaterally with respect to a reference value, and bilaterally with respect to each other. The understanding of the relationship between the *measurements* and the *measurand* determines how the former should be combined to produce an estimate of the latter, and how the uncertainty of this estimate should be assessed. This understanding is best expressed by means of a *statistical model*

(that is, an *observation equation* [1]) that describes that relationship precisely, in particular, how the measurement values depend on the measurand. In the context of inter-laboratory studies and key comparisons, this suggests how the measurement results from the participating laboratories should be combined, and how other, pre-existing information about the measurand should be blended in.

Typically, in an inter-laboratory study, the participating laboratories provide measurements $x_1,...,x_n$, their standard uncertainties $u_1,...,u_n$, and possibly the accompanying degrees of freedom $\nu_1,...,\nu_n$. Each laboratory's measurement summarizes replicated measurements, each of which involves the combination of indications and measured values of participating quantities (thermal expansion coefficient, temperature, etc.), and possibly also other pre-existing information about the measurand. The standard uncertainties are computed using methodology based on the *Guide to the Expression of Uncertainty in Measurement* (GUM) [2], combining uncertainty components from various sources, evaluated either by Type A or Type B methods. The current practice in most metrological experiments is to produce a reference value (x_{ref}) to estimate the measurand and then, in key comparison experiments, to compute the unilateral, $x_i - x_{ref}$, and bilateral $x_i - x_j$ degrees of equivalence (DoE). The reference value (RV) in a key comparison is abbreviated as KCRV.

It is often the case that when a reference value and its uncertainty are computed, plots and other summaries of the data suggest that the measurements are not consistent, that is, that some of the laboratories' measurements are too far away from the reference value with respect to the accompanying uncertainties. In key comparisons, there have been proposals for formal methods to address this problem. Namely, [3,4] suggest using a hypothesis test based on a chi-square statistic to determine whether the participating laboratories belong to a consistent set. If the test determines this not to be the case, then a method is proposed for finding a consistent subset of the laboratories. The KCRV is then computed using only data from laboratories in the consistent subset. Even though not formally expressed in the literature, the chi-square statistic used in this method requires an assumption of a particular statistical model for the measurements.

In this paper we address this issue, and propose alternative models for the analysis of inter-laboratory comparisons that do not require the preliminary identification of a "consistent" subset of laboratories whose measurements are combined into the KCRV, all the others being excluded. Our models are probabilistic: we use probability distributions to describe incomplete knowledge,

and to describe also the dispersion of values that arise in replicated measurements. And our usage of them is statistical: we employ principles of inference to combine information (from data, and possibly from other sources) to produce estimates of the measurand, and to characterize the uncertainty of these estimates. We illustrate our methods on data from key comparison CCL-K1, calibration of gauge blocks by interferometry [5,6]. Section 2 introduces the current recommended methods, the common estimates of reference values, their uncertainties, degrees of equivalence and the consistency testing procedure. Section 3 presents proposed methodology based on the laboratory effects model, gives formulas and interpretations, and an example. A Laboratory Effects Model for experiments with multiple related measurands is shown in section 4.

2. Current practice

2.1. *Methods*

Given the measurements and the standard uncertainties for a particular measurand from each of *n* laboratories, the two most common reference values are the arithmetic average and the weighted mean. These are frequently used statistical estimators of population means that are optimal under certain assumptions about the data. The arithmetic average $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ would be optimal if the measurements were like outcomes of independent, Gaussian random variables, all with the same mean μ, and the same variance σ^2. The weighted mean $\bar{x}_w = \sum_{i=1}^{n} x_i / u_i^2 \Big/ \sum_{i=1}^{n} 1/u_i^2$ would be optimal if measurements were like outcomes of independent (assumption 1), Gaussian (assumption 2) random variables, all with the same (assumption 3) mean μ, and different known variances given by $u_1^2,...,u_n^2$ (assumption 4). We will refer to this as Model W. In most cases the standard uncertainties cannot reasonably be regarded as though they were based on infinitely many degrees of freedom. In some cases, including that of our example CCL-K1, the number of degrees of freedom associated with the standard uncertainties is available. In these circumstances it is preferable to fit yet another model (WD), which relies on the validity of assumptions 1-3, but not on 4. Under this model, the u_i^2 are estimates of the true variances σ_i^2 and behave like outcomes of chi-squared random variables: specifically, $\nu_i u_i^2 \Big/ \sigma_i^2$ is like an outcome of a chi-square random variable with ν_i degrees of freedom.

The weighted mean (Model W) is suggested as the preferred method for calculating a RV by [3] with the condition that a consistency test is carried out using the statistic $\chi^2_{Obs} = \sum_{i=1}^{n} \frac{\left(x_i - \overline{x}_w\right)^2}{u_i^2}$ [4]. Under the assumptions of Model W, this is an observed value of a random variable with chi-square distribution with n-1 degrees of freedom. The consistency test checks whether $\chi^2_{Obs} > \chi^2_{\alpha,n-1}$, the 100(1 - α) percentile of the chi-square distribution with n-1 degrees of freedom. If the χ^2_{Obs} is in fact larger than $\chi^2_{\alpha,n-1}$, the measurements are determined to be inconsistent. The largest subset of participants that passes this test is called the largest consistent subset (LCS) in [4] and is used to compute the reference value. This subset is found using a sequential procedure, computing χ^2_{Obs} for the larger subsets first.

Such consistency testing suffers from several consequential shortcomings. Most importantly, a conclusion that a set of laboratories is inconsistent appears to be interpreted as a violation of assumption 3. This is not necessarily so, in fact it could equally well reflect a violation of any of the other assumptions: particularly assumption 4, which would mean that the uncertainties are underestimated. (Also see [5]) Secondly, the proposed choice of α equal to 0.05 is conventional but arbitrary. This is unfortunate because the LCS procedure may result in different subsets for different values of α as is later shown for the CCL-K1 data. Matters are further complicated by the fact that the proposed procedure typically comprises a possibly large number of component tests, each of which has a probability 0.05 of erroneously rejecting consistency, by chance alone 5% of the subsets will be deemed inconsistent when in fact they are not: the value of α must be adjusted to compensate for this effect of multiplicity.

Another issue derives from the sequential nature of the procedure. Suppose that $\chi^2_{Obs} > \chi^2_{\alpha,n-1}$, and that the hypothesis of consistency was in fact incorrectly rejected. This means that purely by chance, as will occur with probability α, the observed values $x_1, \ldots, x_n$ produce large values of χ^2_{Obs}. Once the whole set of laboratories has been deemed inconsistent, we proceed to test each subset comprising all but one laboratory for consistency in turn, using the corresponding chi-squared statistic. Since the inconsistency of the whole set means that the measurements are over-dispersed to begin with, this second round of testing likely will produce values of the test statistic that are larger than what one would expect if the tests were not being done as a follow-on to the first test. This means that for an α level test, the probability of rejecting the null

hypothesis when it is true is in fact larger than α. Thus using $\chi^2_{Obs_k} > \chi^2_{\alpha,n-2}$ for the follow-up test is not as strong an evidence against the second null hypothesis as is believed, leading to incorrect conclusions.

Next we illustrate these methods using the data from key comparison CCL-K1.

2.2. *Example CCL-K1*

This key comparison was carried out to compare deviations from nominal length of several steel and several tungsten gauge blocks, and is fully described in [6,7]. Here the results for the tungsten gauges are used to illustrate the various procedures. Table 1 has the results for the tungsten 1.1 mm gauge.

Table 1. Deviations from nominal length (in mm), standard uncertainties (in mm), and degrees of freedom for the 1.1 mm tungsten block (20'20987).

Laboratory	Deviation from nominal (mm)	Standard uncertainty (mm)	Degrees of Freedom
OFMET	-54	9	500
NPL	-51	14	119
LNE	-36	10	94
NRC	-51	13	9
NIST	-38	9	50
CENAM	-72	7	72
CSIRO	-32	9	207
NRLM	-66.4	10.3	5
KRISS	-62	9.4	24

Figure 1 shows a plot of the measurements with their 95% confidence intervals and the weighted mean with its uncertainty, based on Model W, and on Model WD. The plot shows possible lab deviations for at least two laboratories (CENAM and CSIRO), and also shows that the measurements from several others are quite far from the weighted mean. If those deviations do not merely reflect a spuriously bad day in the lab, but indeed express the natural interlaboratory variability, then neither Model W nor Model WD would be good choices for this particular set of data.

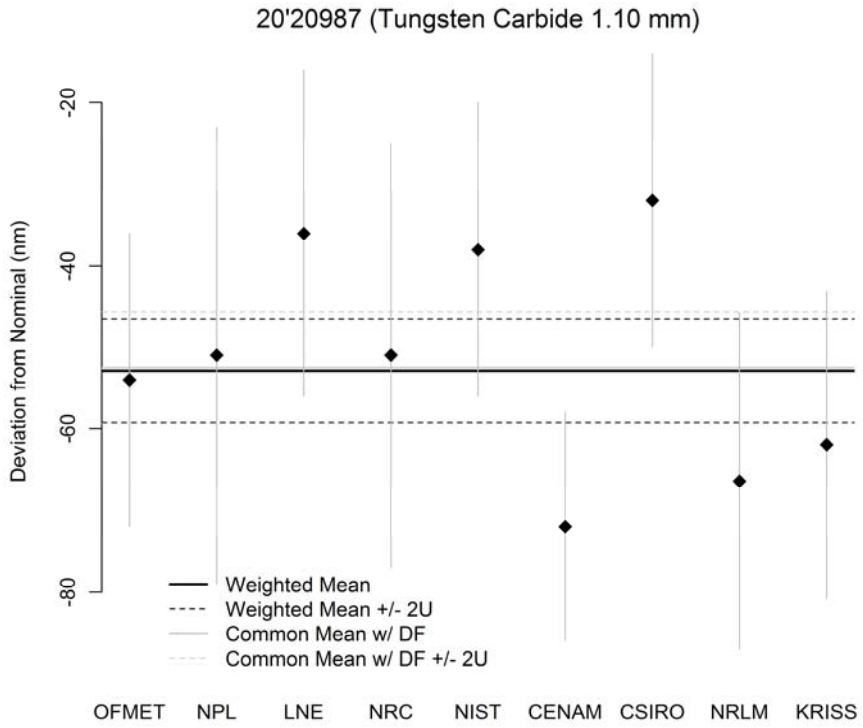

Figure 1. Laboratory measurements with their 95% confidence intervals, the weighted mean with its uncertainty, based on Model W, and on Model WD, for the 1.1 mm tungsten gauge.

A formal test of consistency using the chi-square statistic produces $\chi^2_{Obs} = 21.15$, which is larger than $\chi^2_{0.05,8} = 15.5$, thus confirming what the plot suggests, rejecting the full set of laboratories as consistent at this level of α. The LCS procedure for $\alpha = 0.05$ identifies CENAM as an outlier. The remaining laboratories form the LCS.

Since the level of the test is arbitrary, it is interesting to see what happens for other values of α. Figure 2 contains a plot which illustrates this.

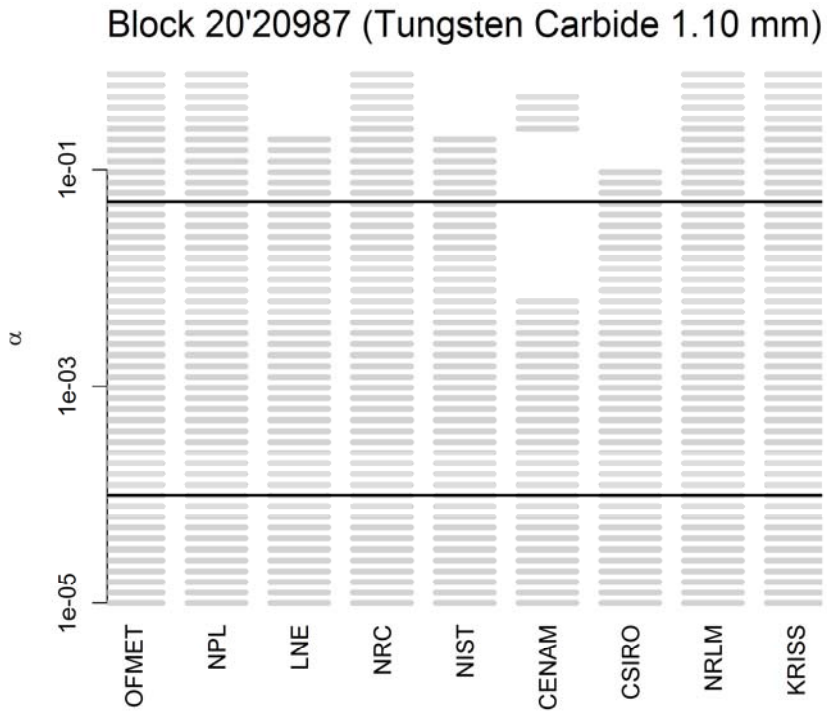

Figure 2. Consistent subsets as α varies. The top horizontal line is for $\alpha = 0.05$. The continuous horizontal line towards the bottom is the value of α that results from the so-called Bonferroni [7, Section 10.7] correction to accommodate the fact that, to find the largest consistent subset, one is either implicitly or explicitly performing $2^n - n - 1$ separate (and statistically dependent) consistency tests. Since α is the probability of incorrectly concluding that the measurements are mutually inconsistent, for sufficiently small α no lab is left out of the largest consistent subset (LCS). As α increases, the lab (CENAM) with the most deviant measurement drops out from the LCS. However, for even larger values of α (somewhat above 0.1), CENAM reenters the LCS, while LNE, NIST, and CSIRO drop out. This illustrates the non-monotonic, counter-intuitive behavior of the LCS as a function of α.

This figure shows that the largest consistent subset (LCS) is not a monotonic function of α: in other words, the LCS corresponding to a particular value of α does not necessarily include the LCS corresponding to a larger value of α. Since any choice of a value for α is both arbitrary and influential, and the corresponding chi-squared test that is employed to judge significance (either using the chi-squared reference distribution or its counterpart determined by simulation) typically has low power to detect heterogeneity, we believe that basing the KCRV on the LCS as [9] recommends generally is imprudent. Indeed, we much prefer alternatives where the measurements provided by all the

laboratories are allowed to speak, and play a role, albeit one that is modulated by their intrinsic uncertainties, and by the extrinsic dispersion of the measurements.

3. Laboratory Effects Model

3.1. *Methods*

Under the Laboratory Effects Model (LEM), measurements are modeled as outcomes of independent Gaussian random variables with means λ_i, and known variances given by $u_1^2,...,u_n^2$. (A similar relaxation of the assumption on the variances as given by Model WD can also be used here.) The means can be written as $\lambda_i = \mu + \beta_i$, where μ is the measurand. There are two well-known [10,11] alternative models which differ in the definition and interpretation of the β_i.

The first type of LEM is called the Fixed Effects Model [10], which assumes that the β_i are systematic laboratory effects (biases) which are expected to re-occur in repeated similar experiments. They are unknown constants to be estimated from the measurement data. In terms of an observation equation, this model can be written as

$$\begin{aligned} X_i &= \mu + \beta_i + E_i \\ E_i &\sim N\left(0, u_i^2\right) \end{aligned} \quad . \tag{1}$$

If the measurand is known independently of the results of the interlaboratory study – for example, it pertains to a standard reference material (SRM) whose certificate puts it at *M*, with uncertainty *u(M)*, then the β_i can be estimated as $x_i - M$ with uncertainty $u_i + u\left(M\right)$. Clearly, the unilateral DoE is $x_i - M$, the bilateral DoE the usual $x_i - x_j$.

However, in most cases no independent information about the measurand exists and the reference value needs to be estimated from the measurements. In such a case, without any additional constraints, unique estimators of μ and of the β_i do not exist [10, section 6.2]. Without some additional (or prior) information about the laboratory biases, the constraint $\sum_{i=1}^{n} \beta_i = 0$ is sensible and leads to unique estimators via least squares or maximum likelihood. The estimator of the measurand μ is $\bar{x}$, the β_i are estimated by $x_i - \bar{x}$. It is a particularly pertinent feature of this model that the estimates of the laboratory effects are in fact the

unilateral DoE with respect to the arithmetic average [12]. The uncertainty of the $\bar{x}$ is

$$u(\bar{x}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} u_i^2} \ , \tag{2}$$

the uncertainty of the unilateral DoE for the ith laboratory is

$$u_i(DOE) = \sqrt{u_i^2 + \frac{1}{n^2}\sum_{j=1}^{n} u_j^2 - \frac{2}{n}u_i^2} \ . \tag{3}$$

To summarize, the Fixed Effects Model does not require a common mean but partially attributes differences among measurements from various laboratories to differences among the means of the random variables which represent the observations. The laboratory uncertainties are not increased in this model. In addition to the classical interpretation of the measurements as observations of random variables, or as sample averages, one may also interpret the measurements as expected values of belief distributions about the measurand. In this case, the u_i are interpreted as standard deviations of such belief distributions. Both interpretations lead to the same formulas and results with different interpretations of the uncertainties and uncertainty intervals. Refer to [12] for a full discussion.

The second type of LEM is called the Random Effects Model [10, 11]. Here the β_i are random biases which are *not* expected to have the *same* value on repeated similar experiments, but instead are outcomes of Gaussian random variables with mean 0 and variance σ_β^2. In terms of an observation equation this is

$$\begin{aligned} X_i &= \mu + \beta_i + E_i \\ \beta_i &\sim N(0, \sigma_\beta^2) \\ E_i &\sim N(0, u_i^2) \end{aligned} \tag{4}$$

Measurements from all laboratories have the same mean μ, but the variances are inflated to $u_i^2 + \sigma_\beta^2$. Using the method of maximum likelihood [11], all of the parameters of this model can be estimated. As the laboratory effects are

random variables, their estimates are really estimates of the realized values of the β_i. The estimated covariance matrix provides the uncertainties of the $\hat{\mu}$ (KCRV) and of the $\hat{\beta}_i$. The $\hat{\beta}_i$ can be used to compare laboratories to each other instead of the standard DoE.

To summarize, under the Random Effects Model, apparent differences among laboratory measurements are explained by an increase in the laboratory uncertainties. The measurements from all laboratories have the same mean. As the ratio of $\hat{\sigma}_\beta^2$ to the u_i^2 gets large for most or all of the laboratories, the $\hat{\beta}_i$ approach $x_i - \hat{\mu}$, the usual unilateral DoE. On the other hand, as this ratio approaches zero, the $\hat{\beta}_i$ approach 0.

As both models are designed to fit data from experiments such as key comparisons and inter-laboratory studies, it is useful to discuss when to apply one versus the other.

The Fixed Effects Model is meant to be used when the laboratory effects are essentially constant biases which are likely to re-occur in similar sizes both relative to the value of the measurand and relative to each other, across similar experiments.

The Random Effects Model is used for situations when there is no such expectation, or no solid documentation for such biases. That is, the laboratory biases are most likely due to some common underlying cause which acts in a *varying nature* so that the laboratories may be closer or farther from the target value in different experiments, purely by chance. There are no identifiable causes that seem responsible for these laboratory biases, hence they are treated symmetrically (relative to each other), and modeled as random variables with a common distribution, centered at 0, and whose variance captures the dispersion of the lab measurements in excess of what their respective uncertainties suggest such dispersion should be.

In each particular experiment, it may be difficult to discern the cause of the variability among laboratories and thus decide which model to use. But it will be seen in the following section that the Random Effects Model provides the most conservative solution.

3.2. *Example*

The laboratory effects models are now applied to the data from key comparison CCL-K1. Table 2 contains the estimates of μ and of the corresponding uncertainty under the three models, Model W, the Fixed Effect Model and the

Random Effects Model. All model-fitting was done in the R [13] environment for statistical computation and graphics, in particular by employing the R function ***lme*** [14], which fits linear Gaussian models possibly containing both random and fixed effects, as [15] describes in detail.

Table 2. Estimates of the mean deviation from nominal length (in mm), standard uncertainties (in mm) in parentheses, and estimates of the standard deviation of the laboratory effects distribution. All under the three alternative models.

Block	**Model W**	**Fixed Effects**	**Random Effects**	$\hat{\sigma}_\beta$
0.5	22 (3.1)	24 (3.5)	24 (4.0)	12
1.00	14 (3.1)	16 (3.5)	16 (4.0)	12
1.01	26 (3.1)	28 (3.5)	27 (3.6)	9
1.10	-53 (3.2)	-51 (3.5)	-51 (4.4)	13
6.0	-48 (3.1)	-47 (3.5)	-47 (3.4)	10
7.0	29 (3.1)	30 (3.5)	31 (3.8)	11
8.00	47 (3.1)	49 (3.5)	49 (3.3)	10
80.0	104 (4.0)	104 (4.7)	104 (4.1)	10
100.0	-76 (4.3)	-79 (5.2)	-79 (4.7)	14

The results show that estimates of the mean deviation from nominal length are quite similar under the three models. The uncertainties of the weighted mean are uniformly smaller than the uncertainties of the other two estimates. Figure 3 provides a comparison of the laboratories via their unilateral Degrees of Equivalence under Model W, and under the Fixed Effects Model. The $\hat{\beta}_i$ are given for the Random Effects Model.

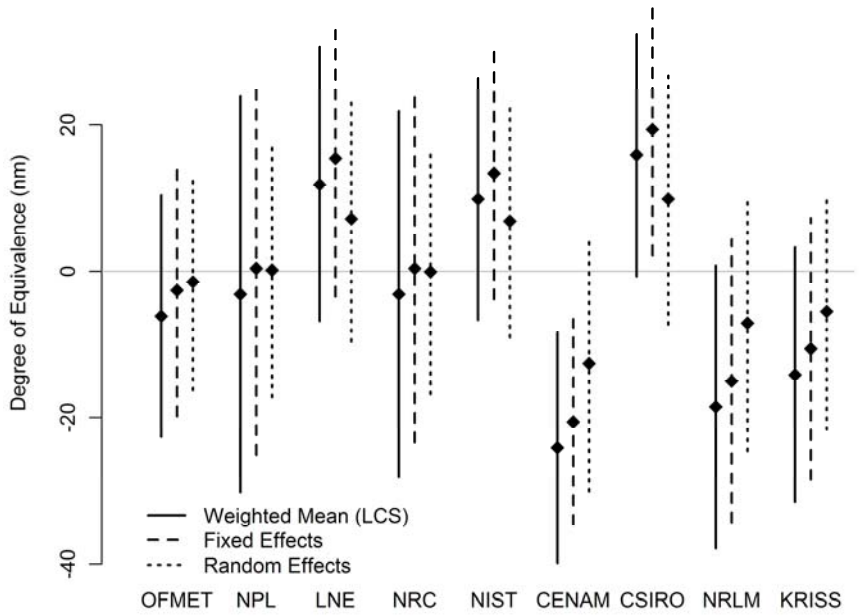

Figure 3. 95% uncertainty intervals for the DoE obtained using the LCS procedure, and using the Fixed Effects model. 95% uncertainty intervals for the estimated biases under the Random Effects model

The plot shows that there are no significant (that is, non-zero) laboratory biases under the Random Effects Model. This happens because this model includes an added variance component to explain the dispersion of the measured values produced by the different laboratories above and beyond the variability that their stated uncertainties alone would suggest. Model W with the LCS procedure identifies CENAM as not belonging to the largest consistent subset. Thus the added variability in the form of the between-laboratory variance σ_β^2 does explain the dispersion of the laboratory measurements under the assumption of a common mean. The Fixed Effects Model on the other hand produces significant laboratory biases for CENAM and for CSIRO which could be interpreted as significant unresolved deviations. The question then is whether these would recur systematically when the same laboratories measure other measurands using the same methods and procedures. If that is the case then the Fixed Effects Model is appropriate, otherwise the Random Effects Model is preferred. This question of model selection can be answered in situations where the experiment consists of measurements on several measurands as is true for CCL-K1.

4. Model Selection Based on Measurements of Multiple Measurands

In CCL-K1, the participating laboratories measured nine steel gauges and nine tungsten carbide gauges, of rectangular cross-section and different lengths, using optical interferometry, applying the same method of fringe fractions, with phase corrections. (Since neither NIM not VNIIM applied these corrections, we follow [6,7] in leaving their measurements out of this study.) Experiments such as this, where multiple measurands are measured using essentially the same methods, afford a unique opportunity to determine whether the apparent differences between laboratories, indicated by the estimated laboratory biases, indeed remain constant across measurands. Figure 4 shows the average predicted laboratory biases from the Random Effects Model for each of the tungsten gauges with different symbols for each laboratory.

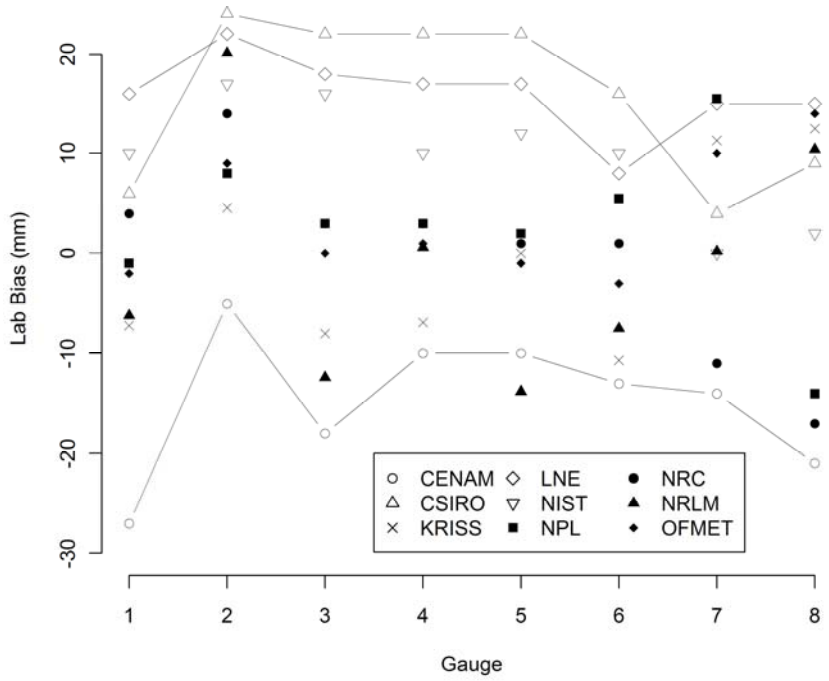

Figure 4. Average Predicted Laboratory Biases in mm.

This graph does show a pattern for the biases of several of the laboratories, particularly LNE and CSIRO tend to be high, and CENAM tends to be low, compared to the other laboratories. The following model can capture such information.

Measurements x_{ij}, where *i* denotes the laboratory and *j* denotes the gauge block, are outcomes of independent Gaussian random variables with means λ_{ij}, and variances given by $u_{11}^2,...,u_{1p},...,u_{n1}^2,...,u_{np}^2$. The number of laboratories is *n* and the number of blocks is *p*. The means can be written as $\lambda_{ij} = \mu_j + \beta_{ij}$, where μ_j are the measurands, and β_{ij} are the laboratory effects. For each laboratory *i*, the β_{ij} are considered to be Gaussian random variables with a mean given by α_i. In terms of an observation equation this is

$$\begin{aligned} X_{ij} &= \mu_j + \beta_{ij} + E_{ij} \\ \beta_{ij} &\sim N\left(\alpha_i, \sigma_{\beta_i}^2\right) \quad . \\ E &\sim N\left(0, u_{ij}^2\right) \end{aligned} \tag{5}$$

Thus the α_i represent the average laboratory bias that applies to all blocks measured by laboratory *i*. As in the simpler, Fixed Effects Model, for a unique solution to exist, there must be a constraint on the laboratory biases. In this model, without any additional knowledge, the constraint $\sum_{i=1}^{n} \alpha_i = 0$ serves this purpose. This model can again be fitted using maximum likelihood methods, in R or other software packages such as WinBUGS [16]. Figure 5 shows a plot of 95% uncertainty intervals for the α_i for the 9 tungsten gauges in the CCL-K1 key comparison. The three laboratories LNE, CSIRO and CENAM do have non-zero average predicted biases over different sized gauges and thus appear to have systematic laboratory biases which can now be interpreted as significant unresolved differences that recur from experiment to experiment. The remaining laboratories do not appear to have such systematic biases as their uncertainty intervals for the α_i include 0.

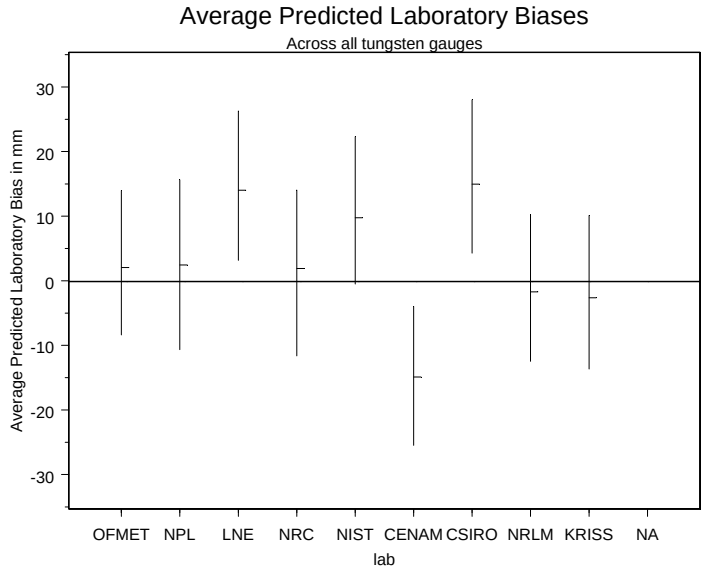

Figure 5. Average estimated laboratory biases α_i (in mm) and their uncertainty intervals, estimated from the measurements of all tungsten gauge blocks

5. Conclusions

The laboratory effects models presented in this paper facilitate the statistical analysis of key comparison data without excluding measurements made by any of the participating laboratories. Such exclusion should be done only for cause, without which even the most discrepant measurement cannot logically be ruled out as erroneous. This inclusive policy enacts a price in uncertainty for the reference value. But this only properly reflects the common state of knowledge that results when the dispersion of the measurements exceeds what one might expect based only on the measurement uncertainties that the laboratories state.

The Fixed Effects Model should be used when there is evidence that the observed deviations express systematic effects of the laboratories' measurement methods and procedures, hence would recur were the intercomparison to be repeated in the same circumstances. Patterns of deviations corresponding to multiple measurements, as the exercise in section 4 illustrates, can provide that evidence. If no evidence is available that incontrovertibly supports the Fixed Effects Model, then the Random Effects Model should be the default choice. As indicated above, in many cases this tends to inflate the KCRV's uncertainty

relative to the individual laboratories' stated uncertainties: however, this merely recognizes the actual dispersion of the measured values.

The Fixed Effects Model allows direct computation of degrees of equivalence as required by the MRA. In the case of the Random Effects Model, where laboratory biases are modeled as random variables with the same probability distribution, equivalence between laboratories may be measured by comparing the estimates of the realized biases.

Acknowledgments

The authors wish to thank Rudolf Thalmann, METAS, for sharing the degrees of freedom for the data from key comparison CCL-K1. The manuscript benefitted greatly by reviews from Gregory F. Strouse and Nien-fan Zhang of NIST.

References

1. Possolo A and Toman B 2007 Assessment of measurement uncertainty via observation equations Metrologia 44 464-475
2. ISO Technical Advisory Group, Working Group 3, Guide to the Expression of Uncertainty in Measurement, International Organization for Standardization, Geneva (1993).
3. Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589 – 95.
4. Cox M G 2007 The evaluation of key comparison data: determining the largest consistent subset *Metrologia* **44** 187-200.
5. Kacker R, Forbes A, Kessel R, Sommer K-D 2008 Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations Metrologia 45, 257 – 265
6. Thalmann R 2001 CCL Key Comparison CCL-K1: Calibration of gauge blocks by interferometry — Final report. Swiss Federal Office of Metrology METAS, Wabern, Switzerland
7. Thalmann R 2002 CCL key comparison: calibration of gauge blocks by interferometry *Metrologia* **39** 165–177
8. Wasserman L 2004 All of Statistics, A Concise Course in Statistical Inference, Springer Science+Business Media, New York, NY, ISBN 0-387-40272-1
9. Decker J, Brown N, Cox M G, Steele A, and Douglas R 2006 Recent recommendations of the Consultative Committee for Length (CCL) regarding strategies for evaluating key comparison data. *Metrologia* 43: L51-L55
10. Searle S R 1971 *Linear Models* John Wiley & Sons New York
11. Searle S R, Casella G, McCulloch C 1992 *Variance Components* John Wiley & Sons New York

12. Toman B 2007 Statistical interpretation of key comparison degrees of equivalence based on distributions of belief *Metrologia* 44 L14-L17
13. R Development Core Team (2008) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org
14. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core team (2008) nlme: Linear and Nonlinear Mixed Effects Models, R package version 3.1-89, http://www.R-project.org
15. Pinheiro J. C., Bates, D. M. (2000) Mixed-Effects Models in S and S-Plus, Springer-Verlag, New York
16. Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. 2000 WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**:325--337.