

# Model Based Uncertainty Analysis in Interlaboratory Studies

Blaza Toman & Antonio Possolo

Statistical Engineering Division  
Information Technology Laboratory  
National Institute of Standards and Technology

June 25, 2008

# Orientation

---

- 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

# Orientation

---

- This understanding is best expressed by means of a *statistical model* (that is, an *observation equation*) that describes that relationship precisely — in particular, how the measurement values depend on the measurand.

# Orientation

---

- In the context of *interlaboratory studies* and key comparisons, this suggests how the measurement results from the participating labs should be combined, and how other, pre-existing information about the measurand should be blended in.

# Outline

---

- Example – Key Comparison CCL-K1
  - Alternative estimates for the reference value
  - Degrees of Equivalence and their uncertainties
  - Heterogeneity of lab measurements and the excess variance problem
  - Shortcomings of consistency testing
  
- Modeling Lab Effects
  - Fixed Effects Model
  - Random Effects Model
  
- Model Selection and Interpretation for Interlab Studies

# CCL-K1

---

## References:

R. Thalmann (2001) . *CCL Key Comparison CCL-K1: Calibration of gauge blocks by interferometry — Final report*. Swiss Federal Office of Metrology METAS, Wabern, Switzerland

R. Thalmann (2002) . CCL key comparison: calibration of gauge blocks by interferometry. *Metrologia* 39: 165–177

# CCL-K1 – Data and Models

---

For each block, measured values  $x_1, \dots, x_n$ , their uncertainties  $u_1, \dots, u_n$ , and degrees of freedom  $\nu_1, \dots, \nu_n$ , from  $n$  labs

Typically, each lab's measurement summarizes replicated measurements (left and right wringing), and each of these in turn involves the combination of indications and measured values of participating quantities (thermal expansion, coefficient, temperature, etc.)

# Modeling Approach

---

- 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

# Interlaboratory Comparison

---

## Objectives:

- Compare laboratories
  - (3) Unilaterally
  - (4) Bilaterally
  
- Compute a Reference Value
  - an estimate of the measurand

# Common Reference Values

- Arithmetic average  $\bar{x} = (x_1 + \dots + x_n)/n$

would be “optimal” if measurements were like outcomes of independent, Gaussian random variables  $X_1, \dots, X_n$ , all with the **same** mean  $\mu$  and variance  $\sigma^2$

Assessment of uncertainty  $u(\bar{x})$  as  $\text{SD}(x_1, \dots, x_n)/\sqrt{n}$  is consistent with that assumption yet disregards  $u_1, \dots, u_n$

# Common Estimates

■ Weighted Average 
$$\bar{x}_w = \frac{(x_1/u_1^2 + \dots + x_n/u_n^2)}{(1/u_1^2 + \dots + 1/u_n^2)}$$

would be “optimal” if measurements were like outcomes of independent, Gaussian random variables  $X_1, \dots, X_n$ , all with the **same** mean  $\mu$  and **different** variances  $u_1^2, \dots, u_n^2$

Assessment of uncertainty as  $1/\sqrt{1/u_1^2 + \dots + 1/u_n^2}$  is consistent with GUM, yet ignores the degrees of freedom

# CCL-K1 Data for 1.1 mm Gauge Block

	x	u	v
1	-54	9	500
2	-51	14	119
3	-36	10	94
4	-51	13	9
5	-38	9	50
6	-72	7	72
7	-82	8	107
8	-32	9	207
9	-66.4	10.3	5
10	-62	9.4	24
11	-50	5.4	67

# Two statistical models which assume a common mean

1. Measurements are outcomes of Gaussian random variables with a **common** mean and variances given by  $u_1^2, \dots, u_n^2$

2. Measurements are outcomes of Gaussian random variables with a **common** mean and variances  $\sigma_1^2, \dots, \sigma_n^2$ , known via  $u_1^2, \dots, u_n^2$   $V_1, \dots, V_n$

, which are based on finite degrees of freedom

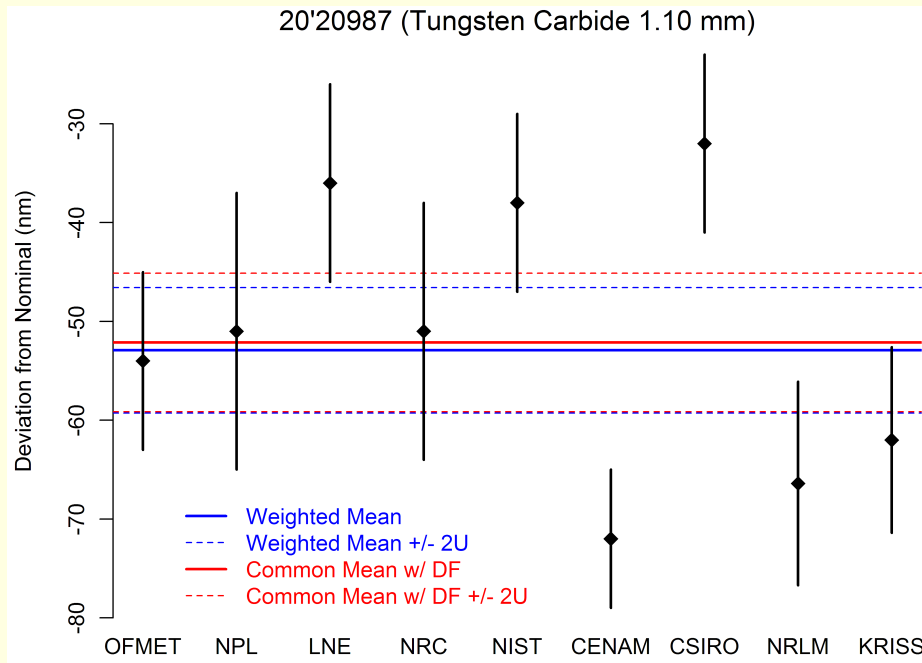
Under these circumstances,  $\frac{V_i u_i^2}{\sigma_i^2}$  is like an outcome of.

a chi-squared random variable with  $V_i$  degrees of freedom

Model 1 underlies the Weighted Mean Analysis

Model 2 is a variation which uses the df, can be fitted using Maximum Likelihood or Bayesian methods

# CCL-K1 Weighted Mean Analysis



The vertical lines are 95% uncertainty intervals based on  $u_i$ .

Tics are the  $x_i$ .

Question the assumptions of both models .

Laboratories CENAM and CSIRO appear to be outliers.

This is too many to be by chance for such a small number of laboratories.

# Testing the assumptions formally - Consistency Testing

■ Compute  $\chi^2_{Obs} = \sum_{i=1}^n \frac{(x_i - \bar{x}_W)^2}{u_i^2}$

Under Model 1, this is an observed value of a random variable with chi-square distribution with  $n-1$  degrees of freedom. Compare this value to  $\chi^2_{\alpha, n-1}$ .

Under Model 2, this chi-square distribution is no longer appropriate, we may then resort to simulation.

# CCLK-1, 1.1 mm gauge block

$\chi^2_{Obs} = 21.15$  is larger than  $\chi^2_{0.05,8} = 15.5$

Possibly (at the 0.05 significance level)

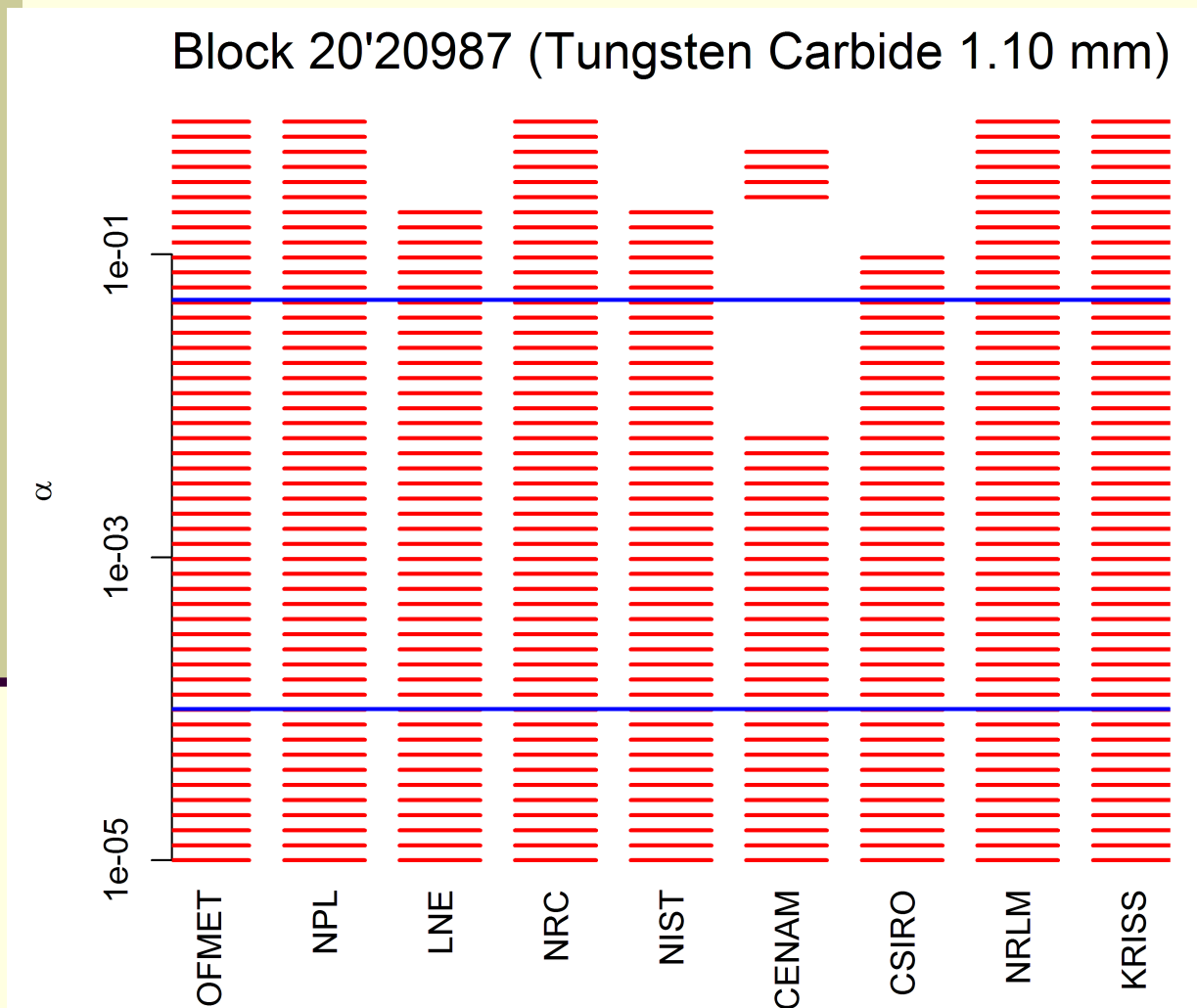
- (a) the common mean assumption is false
- or
- (b) the variances are underestimated
- or
- (c) both (a) and (b)

# Suggested solution to the “no common mean” problem

---

- Cox(2007) - Determine the “Largest Consistent Subset”. Compute RV as a weighted mean of member laboratories’ results.
- This method finds a group of laboratories for which the chi-square test **does not reject** the null hypothesis
  - $H_0$  : Model 1 assumptions hold.
- Potential problems:
  - (1) arbitrary  $\alpha$  level
  - (2) **not reject** does not mean **accept**, the power of the test may be so low that it almost never rejects

# LCS membership as a function of $\alpha$



At  $\alpha = 0.05$ , CENAM is out of the consistent subset.

For higher  $\alpha$ , CENAM will be back in and LNE, NIST and CSIRO will be out.

# Modeling the extra between laboratory variability: Laboratory Effects Model

- Measurements are approximately like outcomes of Gaussian random variables with means  $\lambda_i$ , and variances given by  $u_1^2, \dots, u_n^2$
- The means can be written as  $\lambda_i = \mu + \beta_i$  where  $\mu$  is the measurand.

Observation Equation:  $X_i = \mu + \beta_i + E_i$   
where  $E_i \sim N(0, u_i^2)$

The Gaussian assumption is not critical, under some conditions measurements can be modeled as outcomes of student t random variables or other appropriate distributions.

# Focus on adjustment of the means –

## Fixed effects model

- $\beta_i$  are **systematic** effects (biases) which are expected to recur if similar measurements are made using the same equipment and methods
  - If the measurand is well known( and the principal purpose of the study is to compare the labs), independently estimated by  $M$ , with uncertainty  $u(M)$ , then the  $\beta_i$  can be estimated as  $x_i - M$  with uncertainty  $u(M) + u_i$
  - If no such information exists, the  $\beta_i$  need to be constrained to obtain a solution.

Under  $\sum_{i=1}^n \beta_i = 0$ ,  $\mu$  is estimated by  $\bar{x}$ , and the  $\beta_i$  are estimated by  $DoE_i = x_i - \bar{x}$ , the unilateral Degrees of Equivalence.

Observation Equation:  $X_i = \mu + \beta_i + E_i$   
where  $\sum_{i=1}^n \beta_i = 0$   $E_i \sim N(0, u_i^2)$

# Fixed Effects Model- Uncertainties

■ KCRV: 
$$u(\bar{x}) = \frac{1}{n} \sqrt{\sum_{i=1}^n u_i^2}$$

■ Unilateral DoE: 
$$u(DoE_i) = \sqrt{u_i^2 + \frac{1}{n^2} \sum_{j=1}^n u_j^2 - \frac{2}{n} u_i^2}$$

# Fixed Effects Model- Summary

---

- The apparent differences among lab measurements are explained using constant laboratory biases
- The uncertainties of the measurements are not increased
- Under the constraint that the biases average to 0, the estimated biases are the unilateral DoEs with respect to the arithmetic average
- Pairwise differences between the biases are the usual bilateral DoE

# Focus on the variances – Random Effects Model

$\beta_i$  are biases due to unknown causes, and there is no way to ascertain that they would recur if similar measurements were made using the same equipment and methods: we treat them as non-observable outcomes of Gaussian random variables with mean 0 and variance  $\sigma_\beta^2$

Observation Equation:  $X_i = \mu + \beta_i + E_i$   
where  $\beta_i \sim N(0, \sigma_\beta^2)$        $E_i \sim N(0, u_i^2)$

- Measurements from all laboratories have the same mean  $\mu$ , but the variances of the observations are inflated with respect to the original uncertainties to  $(u_1^2 + \sigma_\beta^2, \dots, u_n^2 + \sigma_\beta^2)$ .
- The measurand is estimated numerically as  $\bar{x}_R$ , the weighted mean with the weights  $u_i^2 + \hat{\sigma}_\beta^2$ . The  $\hat{\sigma}_\beta^2$  is estimated numerically.
- The laboratory effects  $\beta_i$  are random variables, the expected value of their predictive distribution can be used analogously to the unilateral Degrees of Equivalence.

# Random Effects Model - Formulas

For  $n$  laboratories:

$F$  is  $n \times n$  matrix of 1's,

$R$  is  $n \times n$  diagonal matrix of the  $(u_1^2, \dots, u_n^2)$  ,

$D$  is a  $n \times n$  diagonal matrix of  $\hat{\sigma}_\beta^2$

$H$  is a  $n \times 1$  vector of 1s,

$X$  is  $n \times 1$  vector of the data  $(x_1, \dots, x_n)$

Also, define:  $V = (D + R)^{-1}$

# Random Effects Model - Formulas

■ KCRV: 
$$\bar{x}_r = (H'V^{-1}H)^{-1} (H'V^{-1}X)$$

$$u(\bar{x}_R) = (H'V^{-1}H)^{-1}$$

■ Predicted laboratory biases:

$$[\hat{\beta}_1, \dots, \hat{\beta}_n]' = (R^{-1} + D^{-1})^{-1} R^{-1} X$$

with covariance matrix:  $(R^{-1} + D^{-1})^{-1}$

# Random Effects Model -Summary

- Apparent differences among lab measurements explained by an increase in the lab uncertainties.
- All laboratories have the same mean.
- As  $\frac{\hat{\sigma}_\beta^2}{u_i^2} \rightarrow \infty$ , the predicted lab biases approach  $x_i - \bar{x}_R$ , the unilateral DoE. As  $\frac{\hat{\sigma}_\beta^2}{u_i^2} \rightarrow 0$  the predicted biases approach 0.

# Random Effects Analysis - Computation

---

- Maximum Likelihood Estimation
- Bayesian analysis with non-informative priors on  $\mu$  and  $\sigma_{\beta}$  .
- Can use R or WinBUGS.

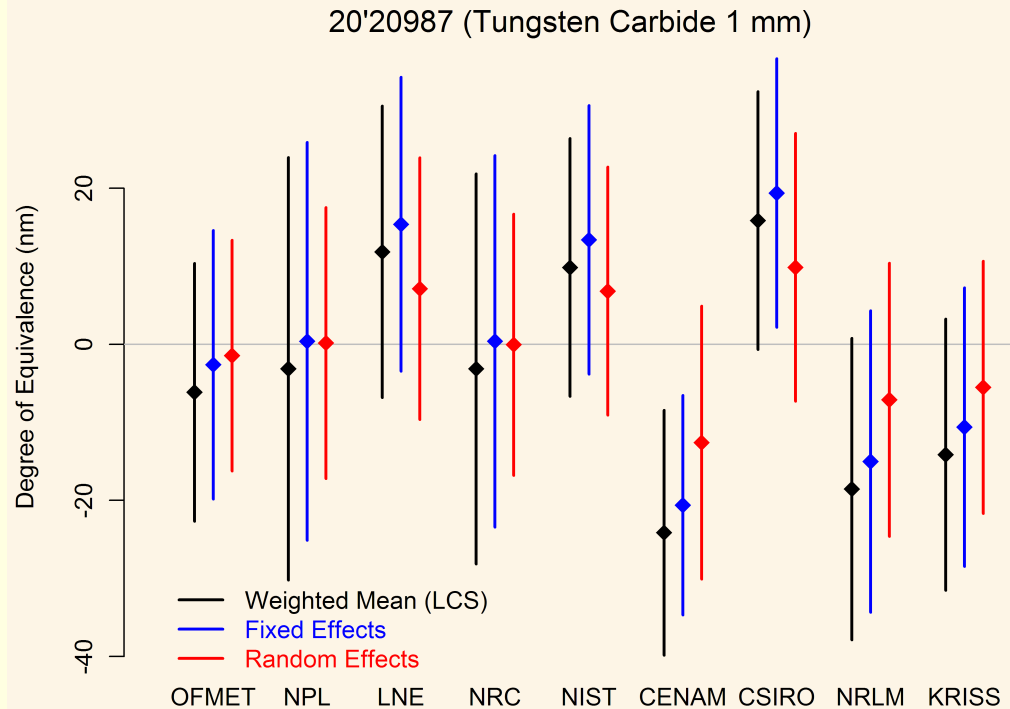
# An Example: Analysis of CCL-K1 KCRV

Block	$\bar{x}_W$	$\bar{x}_R$	$\bar{x}$
0.5	25 (3)	25 (4)	25 (3)
1	14 (3)	14 (5)	15 (3)
1.01	24 (3)	24 (5)	24 (3)
1.1	-55 (3)	-54 (5)	-54(4)
6	-51 (3)	-52 (6)	-51 (4)
7	27 (3)	27 (5)	27 (3)
8	48 (3)	48 (4)	48 (4)
80	103(3)	102(4)	101(4)
100	-75 (4)	-76 (5)	-80 (5)

The estimates of  $\mu$  are similar under Model 1 and the Fixed and Random Effects models.

Estimates of uncertainty are the smallest for Model 1, followed by Fixed Effects, followed by Random Effects model.

# DoEs and predictions of lab biases for 1.1 mm block



Vertical lines are 95% uncertainty intervals for DoEs or 95% prediction intervals for lab biases.

The Random effects model shows fewer significant deviations than the Fixed effects model

# Which model – Fixed or Random Effects?

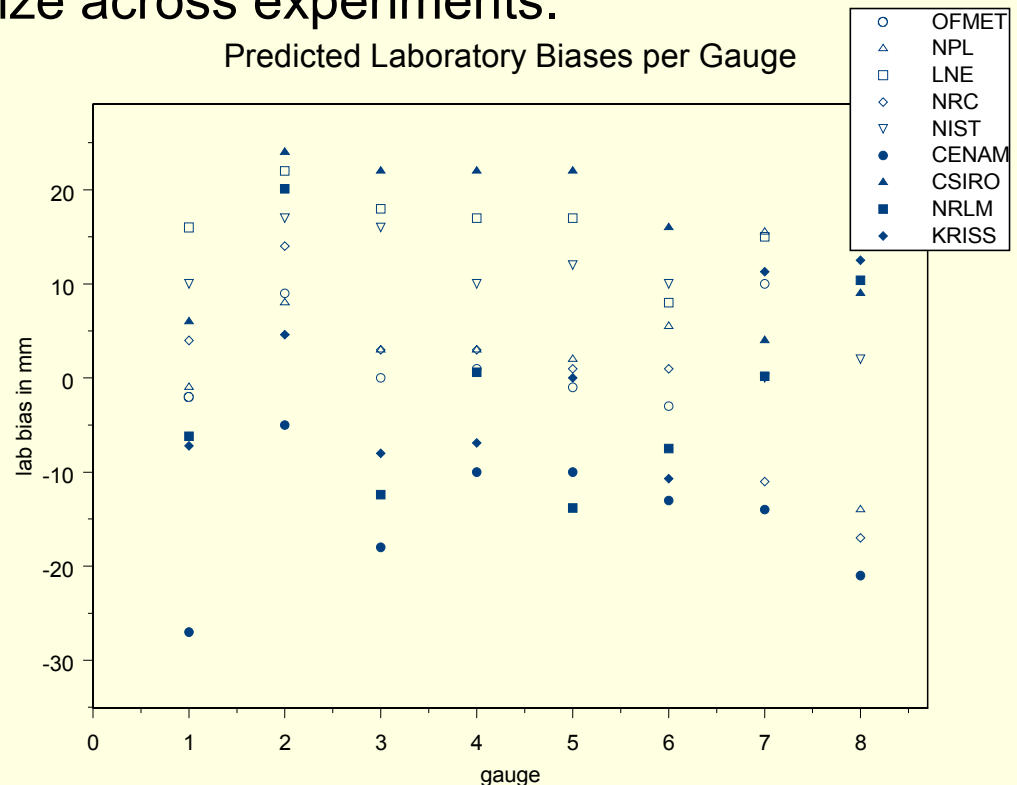
---

- Depends on the particular application.
- If laboratory biases are expected to carry over from experiment to experiment – Fixed Effects.
- If laboratory biases are more of a random nature– Random Effects.
- The Random Effects model is clearly the most conservative in terms of uncertainty.

# Multiple Measurands

When there are multiple ( $p$ ) measurands, it is possible to examine whether laboratory biases are of similar size across experiments.

Example: CCL-K1  
Tungsten gauges



# Laboratory Effects Model for Multiple Measurands

The model: Measurements are outcomes of Gaussian random variables with means ( $\lambda_{ij}$ ) and variances given by  $u_{ij}^2$  such that

$$\lambda_{ij} = \mu_i + \beta_{ij}$$

where  $\beta_{ij}$  are Gaussian random variables with means  $\alpha_j$  with constraint

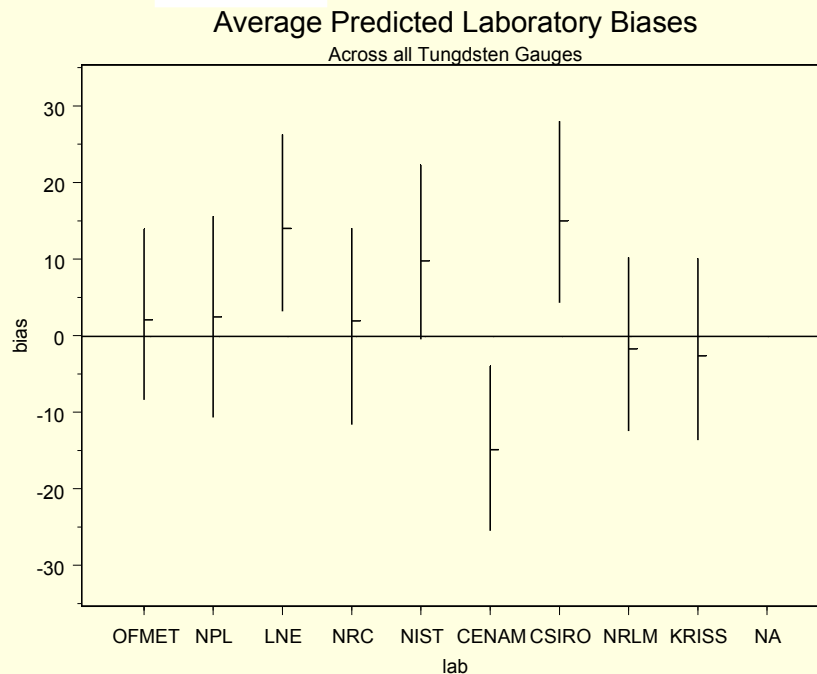
$$\sum_{j=1}^n \alpha_j = 0$$

Observation Equation:

$$\begin{aligned} X_{ij} &= \mu_i + \beta_{ij} + E_{ij} \\ \beta_{ij} &\sim N(\alpha_j, \sigma_{\alpha_j}^2) \quad \sum_{j=1}^n \alpha_j = 0 \\ E_{ij} &\sim N(0, u_{ij}^2) \end{aligned}$$

# Results -

## “average” laboratory biases across gauges



Plot shows estimates (middle horizontal bars), and 95% uncertainty intervals for the  $\alpha_i$ .

Several laboratories show biases that appear to carry over from gauge to gauge.

This provides a method of summarizing Key Comparison data with multiple similar measurands.

# Conclusions

---

- The laboratory effects models facilitate statistical analysis of key comparison data without excluding measurements made by any of the participating laboratories.
- This inclusive policy enacts a price in uncertainty, for the reference value, in some cases, for the degrees of equivalence as well.
- This 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.

# Conclusions - cont.

---

- The fixed effects model allows direct computation of degrees of equivalence as required by the MRA.
- 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 means of the predictive distribution for the biases.
- When the key comparison study includes measurements for multiple measurands measured in similar experiments, it becomes possible to determine whether systematic biases are in fact present.

# Additional Observations and Comments

---

- The proposed approach is a composite of the **top-down** (the laboratory effect model) and **bottom-up** (the laboratories results are products of the measurement model)
- The lab effect model is essentially the model of ISO-5725
- The meaning of uncertainty is consistent with that of the GUM (we think)

# Bibliography

---

- [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] Thalmann R 2001 *CCL Key Comparison CCL-K1: Calibration of gauge blocks by interferometry — Final report*. Swiss Federal Office of Metrology METAS, Wabern, Switzerland
- [6] Thalmann R 2002 CCL key comparison: calibration of gauge blocks by interferometry *Metrologia* **39** 165–177

# Bibliography

---

- [7] Wasserman L 2004 *All of Statistics, A Concise Course in Statistical Inference*, Springer Science+Business Media, New York, NY, ISBN 0-387-40272-1
- [8] 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
- [9] Searle S R 1971 *Linear Models* John Wiley & Sons New York
- [10] Searle S R, Casella G, McCulloch C 1992 *Variance Components* John Wiley & Sons New York
- [11] Toman B 2007 Statistical interpretation of key comparison degrees of equivalence based on distributions of belief *Metrologia* 44 L14-L17