Journal of Data Science **5**(2007), 393-412

Statistics in Metrology: International Key Comparisons and Interlaboratory Studies

Andrew L. Rukhin[1,2] and N. Sedransk[1]
[1] *National Institute of Standards and Technology and*
[2] *University of Maryland at Baltimore County*

Abstract: Stochastic modeling and analysis of international key comparisons (interlaboratory comparisons) pose several fundamental questions for statistical methodology. A key comparison (KC) is specifically designed to derive the key comparison reference value and to assess conformance of calibrations by participating national metrology laboratories at a few, "key", settings for a particular measurement process. An approach to the statistical study of key comparisons data is proposed using a model taken from meta-analysis. This model leads to a class of weighted means estimators for the consensus value and to a method of assessing the uncertainty of the resulting estimates.

Key words: DerSimonian-Laird estimator, interlaboratory study, key comparison reference value, Mandel-Paule algorithm, maximum likelihood, median, meta-analysis, uncertainty of type A, uncertainty of type B, weighted means.

1. Introduction and Summary

The need for statistically sound analytic methods for collaborative metrology studies data motivates this paper. Several approaches are known, but at present there is no commonly accepted methodology for statistical analysis of the interlaboratory studies. The paper begins with a review of the central issues arising in statistical modeling and analysis of international key comparisons data, proceeds to formalize the mathematical structure for these data and then compares several procedures for calculation of the Key Comparison Reference Value (KCRV) and for the estimation of the uncertainty for this value. Evaluation of KCRV is demanded by the main document on international cooperation for measurement quality assurance, the Mutual Recognition Agreement (MRA) (1999) published by Le Comite International des Poids et Mesures. The MRA is realized through KC which typically involve National Metrology Institutes (NMI) each of which analyzes its measurements and reports the results consisting of this NMI's estimate of the measurement value along with the combined standard uncertainty.

Uncertainty, as the term in measurement science or metrology, is distinct from variance, commonly evaluated for other statistical analyses. Uncertainty comprises two components. The first uncertainty component, Type A (aleatory) uncertainty, corresponds directly to estimated standard deviation, and is based upon standard statistical (usually least squares) variance estimates. Thus, Type A errors are either estimable or confirmable from data. However, the second component, Type B (epistemic) uncertainty, draws on expert scientific judgment as well as data and provides information about effects and/or extra-variation that are not observable or are only partially observable within the context of the study itself. In an international KC each participant provides the uncertainties of each type for its measurements. Consequently for combined analysis, it is of critical importance to determine the dependence of each statistical method for calculation of the KCRV and its associated uncertainty upon both Type A and Type B errors in the participants' reports.

Section 2 of this paper examines further Type B uncertainty, and reviews general criteria for statistical estimates of uncertainties. The recently published key comparisons for accelerometers, CCAUV.V-K1 (von Martens, Elster, Link, Taebner and Wabinski, 2002) illustrates key comparisons implementation and the statistical issues involved. Section 3 presents a model for inference inspired by meta-analysis that provides explicit representation of both uncertainties. This model is used to derive a class of weighted means estimators whose properties are presented in Section 4, where approximate confidence intervals are obtained. Section 5 illustrates these methods using the CCAUV.V-K1 study and presents simulation results to compare the estimators. Section 6 summarizes this paper.

2. Principal Issues in Uncertainty Budgets

An International Key Comparison is an interlaboratory study that involves several NMIs, with one of them, the *pilot* laboratory, responsible for coordination of the whole study and for analysis of the combined results from the participants. The purpose of a KC is to facilitate international trade by determining the degree of conformance of measurements made by the various NMIs. For example, based on the KC, a customer for calibration service should be able to assess which of the other NMIs could provide a calibration that would be acceptable for meeting specifications set in his/her own country. (Thus the goal of KC differs from the objective of aggregating experiments to provide the most accurate value for a physical constant with the highest precision possible.) The customer's practical decision requires knowing not only the NMIs' Reference Values and the KCRV, but also uncertainty for each of these.

The design of a KC requires each participating NMI to follow a carefully prescribed measurement process. When preparing the KC measurement protocol, the organizing scientists must rigorously define the individual components to be itemized in calculating Type B uncertainty for that KC. Gleser (1988) describes uncertainty as "measure of spread of the collection of values not rendered *implausible* by the measurement". After stating aleatory uncertainty with data-driven computations, epistemic error then accounts for both systematic effects and random sources of other *plausible* imperfections in a measurement. These could include wear and degradation as well as flaws or idiosyncrasies in the measurement instrument itself; probes used to define locations on a surface have individual imperfectly spherical geometries and they also suffer deformation through use. Input quantities to computed measurands and to correction factors are inexact; environmental measurements such as temperature carry their own uncertainties that depend on the thermometer or other measurement device; measurements involving electron beams are affected by secondary electron scatter. Operators of complex measuring instruments have "off-days"; measurements near the limits of the instrument's range may have peculiar biases or instabilities. Thus resultant epistemic uncertainty can dwarf the more ordinary (data-based and measurable) sources of variation by at least an order of magnitude, especially when the measurement process is ultra-precise. Expert judgment drawing on previous accumulated data when available, on experience or on scientific wisdom, is used to assess epistemic uncertainty according to KC prescription for its components.

Swyt (2001) provides a review of uncertainty, especially Type B, in the context of the evolution of length and dimension measurements. The once international standard for the meter was a platinum-iridium bar with two etches "precisely" one meter apart. There was a parallax problem at the micro-scale since the etched lines themselves have a width. The magnitude of epistemic uncertainty could have been provided by setting an upper bound, based on etching widths, or by invoking expert opinion about the probable divergence due to parallax or by giving a probability distribution over a range of these divergences. Measuring the width of a line is subject to different biases and uncertaintites; no line has perfectly straight edges, neither are the edges irregularitites parallel. Therefore the specific location of measurement contributes to Type B error. However, even in this example epistemic uncertainty is not unique; scientific wisdom is essential as the biases in reading the endpoints of a distance tend to be subtractive when measuring between the lines, and additive when measuring across a solid line.

Now the platinum-iridium bar has been replaced, first by a wavelength of light, later by the propagation of an electromagnetic wave in an interval of time. At this time the length standard is measured in terms of the frequencies of visible light lasers (iodine-stabilized HeNe laser) directly against the cesium-beam atomic clock. With much more complicated measuring instrumentation, every part of the instrument contributes to Type B error and all parts are subject to variation from different sources. Thus the epistemic uncertainty associated with the current standard for the meter is influenced by more than 30 different sources of errors. Some of these are instrument phenomena for which the physics is understood and for which one can derive mathematically limits of variation. Some of the other sources are due to the drift or cyclic behavior of the instrument or to the environmental inputs, and these can be predicted from a long series of observations. Some errors result from variation in inputs to complicated algorithms, and can be modeled by MCMC or other simulations. Still others require expert opinion and a Bayes theory formulation. The Central Limit Theorem motivates normality to represent type B uncertainties encountered in modern KC, and this is the suggested model in Section 3.

In the end, all errors are combined (usually by using a quadrature summation, which tacitly assumes independence.) The final statement of the Type B uncertainty is by the expert who submits itemization by source, and whose final summary value is his/her personal "best judgment" incorporating all available experience and expertise. Upon completion of KC each NMI reports its own measured values (or computed estimates of the Reference Value) and submits its complete Uncertainty Budget consisting of the Type A uncertainty and the Type B uncertainty which are always presented in terms of individual measurements. Indeed a future customer would ordinarily require a single measurement or calibration, but the sample sizes (number of repeats) typically vary among NMI's.

The KC of accelerometers (CCAUV.V-K1, von Martens *et el.*, 2002) was organized to compare measurements of sinusoidal linear accelerometers at specified frequencies in the range from 40Hz to 5kHz. (Each accelerometer measured charge sensitivity at the specified frequencies and at different acceleration amplitudes.) Two types of (single-ended design and back-to-back) accelerometers were employed at each of twelve NMIs (including NIST), with the Physikalish-Technische Bundesanstalt, Germany, serving as the pilot laboratory. Each participating NMI reported its own laboratory means, the within lab variances (Type A uncertainties), and the combined (Type A plus Type B) uncertainties with a complete Uncertainty Budget.

The objective was to determine the KCRV for charge sensitivity and the associated uncertainty (separately for each type of accelerometer and for each specified frequency). Type B uncertainties included errors in accelerometer voltage, amplitude gain, temperature variation, total harmonic distance, translational motion, minimum point resolution, vibrational frequency, displacement quantification, hysteresis, residual interference, etc. . In this heterogeneous study the uncertainties by NMI's differed by factor of 7 to 9 (variances by a factor $50 - 80$).

Beside KCRV with its uncertainty, the statistical analysis of Key Comparison data includes calculation of, for each participating NMI, a degree of equivalence (deviation of NMI value from consensus KCRV) plus combined uncertainty for that deviation. Typically, a laboratory whose measurement results deviate significantly from the KCRV is flagged; the results of these measurements are considered suspicious. Of course, it is possible that such a laboratory provides the best estimate of the measurand, and MRA acknowledges such a possibility.

Thus, a principal issue revolves around the expression of Type B errors and its implications for the KCRV and the degrees of equivalence including their associated uncertainties. In this paper we suggest a hierarchical model for Type B error and derive KCRV estimators. We also give methods to evaluate the variance of these KCRV estimators which provide approximate confidence intervals for KCRV and for the degrees of equivalence. The effects of epistemic uncertainty and its misspecifications on the mean squared errors and coverage probabilities are explored via a Monte-Carlo study.

3. The Meta-Analysis Model

Consider the situation where one the key comparison reference value is to be established by combining information from several, say, p laboratories. We accept a natural model suggested by analysis of variance and meta-analysis. In this model the data in the ith laboratory has an additive error structure consisting of a random laboratory effect λ_i, the laboratory bias b_i (representing type B uncertainty), and the measurement errors ϵ_{ij} (contributing to Type A uncertainty). More precisely, assume that the data Y_{ij} has the form

$$Y_{ij} = \mu + \lambda_i + b_i + \epsilon_{ij}. \tag{3.1}$$

Here $i = 1, \ldots, p$ indexes the laboratories, $j = 1, \ldots, n_i$ with n_i representing the sample size (the number of measurements) in laboratory i; μ is the true mean (the KCRV). The random variables λ_i and ϵ_{ij} are assumed to be mutually independent and normal with zero means and variances σ_L^2 and τ_i^2 respectively.

The random between-laboratories effect λ (interactions after Cochran, 1937, or hidden errors in terminology of Willink, 2002) is often observed in heterogeneous collaborative studies where individual laboratories estimates of the true mean can be quite different. It is possible that in (3.1) $\lambda_i \equiv 0$. The component b_i representing the epistemic uncertainty assessed by the laboratory i is composed of a systematic bias component, δ_i, and a variance component, β_i^2. Define b_i in a hierarchical way, $b_i \mid \delta_i \sim N(\delta_i, \beta_i^2)$, with the expected bias component δ_i for NMI i being normal, $N(\eta, \varphi_i^2)$. Then, marginally, $Eb_i = \eta, Var(b_i) = \varphi_i^2 + \beta_i^2 = \sigma_{B_i}^2$.

We assume that the reported combined standard Type B uncertainty provides an estimate of the variance, $\sigma^2_{B_i}$, whereas the individual estimates of η, φ_i^2 and β_i^2 are not available. Then only the parameter $\theta = \mu + \eta$ (but not μ itself) can be estimated; one of the goals of this paper is to study the effect of the epistemic uncertainty on the performance of confidence intervals for the KCRV μ. A large bias η reduces to substantial deterioration of the existing decision rules, but a large variance component, $\sigma^2_{B_i}$, of the Type B uncertainty is relatively harmless in this regard.

In matrix notation our model can be written as a particular case of the general linear model $\mathbf{y} = \mathbf{X}\theta + \mathbf{u} + e$. Here $\mathbf{y}$ is the total data vector of dimension $n = n_1 + \ldots + n_p$; $\mathbf{X}$ is a $n \times 1$ vector formed by ones, $\mathbf{u}$ is the same size vector formed by p independent, mean zero, random normal components with the variables in the i-th block of size n_i having variance $\sigma_L^2 + \sigma^2_{B_i}$. Thus θ is the unknown parameter (fixed effects) and $\mathbf{u}$ is random effect vector uncorrelated with the errors vector e. Therefore, the equation (3.1) reduces to the classical random effects model with the common mean θ, the between-laboratories effect with the variance σ_L^2 and the errors whose variance is $\sigma^2_{B_i} + \tau_i^2$. A subjective estimate of Type B error, $\sigma^2_{B_i}$ is provided, and statistical estimate of τ_i^2 is also available.

The equation (3.1) leads to the following model for the sample means, $Y_i = \bar{Y}_i = \sum Y_{ij}/n_i$, $Y_i|\mu, \eta \sim N(\mu + \eta, \sigma_L^2 + \sigma_i^2 + \sigma^2_{B_i})$, where $\sigma_i^2 = \tau_i^2/n_i$. Clearly, $EY_i = \theta$, and $E(Y_i - \mu)^2 = \sigma_L^2 + \sigma_i^2 + \sigma^2_{B_i} + \eta^2$. Notice that an unbiased estimate of the variance σ_i^2 (Type A uncertainty) is available via the sample variances, $s_i^2 = \sum_j (Y_{ij} - Y_i)^2/[n_i(n_i - 1)]$ for the i-th laboratory, $i = 1, \ldots, p$. In contrast, $\sigma^2_{B_i}$ (a component of Type B uncertainty) is not directly estimable from the data. In a different model (e.g. Hall and Willink, 2001) the epistemic uncertainty is taken to be the uniform (or triangular) distribution on an interval. However, in the authors' experience based on many interlaboratory studies, this does not present a very realistic model. Besides numerical differences between these distributions are minor.

In the next Section we review several estimators of θ and the estimates of their variance. The study of the classical maximum likelihood estimate of θ in the model (3.1) when $b_i \equiv 0$ was initiated by Cochran (1937) (see Rao, 1981 for a historic review). The solution of the likelihood equations in the situation when $\sigma_i^2 \equiv \sigma^2$ was explored by Harville (1977). This, as well as the solutions for the restricted likelihood, are discussed in Searle, Casella and McCulloch (1992, chap. 6 and 8).

4. Weighted Means Statistics: Approximate Distributions and Confidence Intervals

In our problem the maximum likelihood estimator does not have an explicit form; the negative log-likelihood function (to be minimized) for given Y_i and $s_i^2, i = 1, \dots, p$ is

$$\sum_i \left[\frac{(Y_i - \theta)^2}{\sigma_L^2 + \sigma_{B_i}^2 + \sigma_i^2} + \log(\sigma_L^2 + \sigma_{B_i}^2 + \sigma_i^2) \right] + \sum_i (n_i - 1) \left[\frac{s_i^2}{\sigma_i^2} + \log \sigma_i^2 \right].$$

This function may not be unimodal, but the minimizer in θ has the form, $\hat{\theta} = \sum_{i=1}^p \hat{\omega}_i Y_i$, with the weights $\hat{\omega}_i = (\hat{\sigma}_L^2 + \sigma_{B_i}^2 + \hat{\sigma}_i^2)^{-1} / \sum_j [\hat{\sigma}_L^2 + \sigma_{B_j}^2 + \hat{\sigma}_j^2]^{-1}$. A numerical iterative procedure to find $\hat{\sigma}_i^2$ and $\hat{\sigma}_L^2$ when $\sigma_{B_i}^2 \equiv 0$ is given in Vangel and Rukhin (1999).

Because of the complicated nature of the maximum likelihood estimator simpler procedures are desired in metrological applications. The goal here is to introduce certain weighted means statistics useful for the KCRV estimation. To derive them we assume that $\eta = 0$, so that $Y_i \sim N(\mu, \sigma_L^2 + \sigma_{B_i}^2 + \sigma_i^2)$. Let us start with the case when all variances σ_L^2, σ_i^2 and $\sigma_{B_i}^2$ are known. Then the best unbiased estimator of the reference value μ is a weighted means statistic,

$$\hat{\mu} = \frac{\sum_1^p w_i Y_i}{\sum_1^p w_i}, \tag{4.1}$$

where $w_i = w_i^0 = (\sigma_L^2 + \sigma_{B_i}^2 + \sigma_i^2)^{-1}$. In this situation it is also the maximum likelihood estimator, and without normality assumption (but when all variances are known) this is the best linear (in Y_i) unbiased estimator of μ.

Assume that the reported type B uncertainty, $\sigma_{B_i}^2$, provides a good estimate of the corresponding variance component, and the within-laboratories variances σ_i^2 can be estimated by the available statistics s_i^2. Even without the normality assumption, for arbitrary non-random weights w_i,

$$E \sum_1^p w_i (Y_i - \hat{\mu})^2 = \sum_1^p w_i Var(Y_i) - \frac{\sum_1^p w_i^2 Var(Y_i)}{\sum_1^p w_i}. \tag{4.2}$$

(See Appendix for the proof.) In particular, when $w_i = 1/(\sigma_{B_i}^2 + \sigma_i^2)$, (4.2) gives

$$E \sum_1^p \frac{(Y_i - \hat{\mu})^2}{\sigma_{B_i}^2 + \sigma_i^2} = p - 1 + \sigma_L^2 \left[\sum_1^p \frac{1}{\sigma_{B_i}^2 + \sigma_i^2} - \frac{\sum_1^p \frac{1}{(\sigma_{B_i}^2 + \sigma_i^2)^2}}{\sum_1^p \frac{1}{\sigma_{B_i}^2 + \sigma_i^2}} \right]. \tag{4.3}$$

The idea behind the method of moments suggests the following procedure to estimate σ_L^2 by using (4.3) while replacing σ_i^2 by s_i^2. In the formula for the weights of the form,

$$w_i = \frac{1}{z + \sigma_{B_i}^2 + s_i^2}, \tag{4.4}$$

take a non-negative $z = z_{DL}$, determined as

$$z_{DL} = \max\left[0, \frac{\sum_1^p \frac{(Y_i - \hat{\mu}_{GD})^2}{\sigma_{B_i}^2 + s_i^2} - p + 1}{\sum_1^p \frac{1}{\sigma_{B_i}^2 + s_i^2} - \sum_1^p \frac{1}{(\sigma_{B_i}^2 + s_i^2)^2}\left[\sum_1^p \frac{1}{\sigma_{B_i}^2 + s_i^2}\right]^{-1}}\right],$$

where

$$\hat{\mu}_{GD} = \frac{\sum_1^p \frac{Y_i}{\sigma_{B_i}^2 + s_i^2}}{\sum_1^p \frac{1}{\sigma_{B_i}^2 + s_i^2}}. \tag{4.5}$$

This procedure is a direct extension of the method suggested by DerSimonian and Laird (1986) when $\sigma_{B_i}^2 \equiv 0$. In this case (4.5) is the well-known Graybill-Deal estimator of the common mean. Thus, the statistic $\hat{\mu}_{GD}$ and the weights $(\sigma_{B_i}^2 + s_i^2)^{-1}$ corresponding to $\sigma_L^2 = 0$, are used to evaluate the sum in the left-hand side of (4.3) which is then employed to estimate the unknown σ_L^2. The resulting estimator for μ has the form

$$\hat{\mu}_{DL} = \frac{\sum_1^p \frac{Y_i}{z_{DL} + \sigma_{B_i}^2 + s_i^2}}{\sum_1^p \frac{1}{z_{DL} + \sigma_{B_i}^2 + s_i^2}}. \tag{4.6}$$

A similar extension of the Mandel-Paule algorithm (1982) also uses weights of the form (4.4) in the formula (4.1) for the weighted means statistic. However, now the value, z (designed to estimate σ_L^2) is motivated by the formula which follows from (4.2) when the weights w_i are optimal, i.e., when they coincide with w_i^0,

$$E \sum_1^p w_i^0 (Y_i - \hat{\mu})^2 = p - 1.$$

Thus the Mandel-Paule estimating equation for z is,

$$\sum_1^p \frac{(Y_i - \hat{\mu})^2}{z + \sigma_{B_i}^2 + s_i^2} = p - 1.$$

The explicit solution of this equation for $p \geq 3$ does not exist, in practice a number of iterations is needed to get it with desired accuracy. The following approximation is the one step application of the Newton method for the initial value $z = z_{DL}$. It is based on the formula for the derivative of the weighted sum of squares (e.g. Rukhin, Biggerstaff and Vangel, 2000, p 323) and is easily computable,

$$z_{MP} = \max\left[0, z_{DL} + \frac{\sum_i (Y_i - \hat{\mu}_{DL})^2/(z_{DL} + \sigma_{B_i}^2 + s_i^2) - p + 1}{\sum_i (Y_i - \hat{\mu}_{DL})^2/(z_{DL} + \sigma_{B_i}^2 + s_i^2)^2}\right].$$

The resulting estimator of μ has the form

$$\hat{\mu}_{MP} = \frac{\sum_1^p \frac{Y_i}{z_{MP}+\sigma_{B_i}^2+s_i^2}}{\sum_1^p \frac{1}{z_{MP}+\sigma_{B_i}^2+s_i^2}}. \tag{4.7}$$

Rukhin (2002) reviews the Mandel-Paule estimator and the DerSimonian-Laird procedure when $\sigma_{B_i}^2 \equiv 0$. Notice that $(p-1)^{-1}\sum_1^p (Y_i - \hat{\mu})^2/(\sigma_{B_i}^2 + \sigma_i^2)$ is the square of the so-called Birge ratio which is commonly used in metrology for testing goodness-of-fit (Mohr and Taylor, 1999). Thus, the Mandel-Paule procedure seeks the weights under which the squared Birge ratio equals its expected value.

When the within-laboratories variances σ_i^2 can be assumed to be known (in practice they are taken to be s_i^2), the maximum likelihood estimation of μ was investigated by Willink (2002). This estimator also is a weighted means statistic (4.1) with the weights of the form (4.4) where $z = z_W$ is the minimizer in z of the negative loglikelihood function,

$$z_W = \arg\min_z \sum_1^p \left[\frac{(Y_i - \hat{\mu})^2}{z + \sigma_{B_i}^2 + s_i^2} + \log(z + \sigma_{B_i}^2 + s_i^2)\right]. \tag{4.8}$$

The resulting likelihood equation,

$$\sum_1^p \frac{(Y_i - \hat{\mu})^2}{(z + \sigma_{B_i}^2 + s_i^2)^2} = \sum_1^p \frac{1}{z + \sigma_{B_i}^2 + s_i^2},$$

can be solved iteratively.

For all these statistics, the weights have the form (4.4). To estimate the variance of $\hat{\mu}$, the following procedure suggested in a more general setting of linear models by Horn, Horn and Duncan (1975) provides a good answer. Let $\omega_i = w_i/(\sum_1^p w_k)$, $\sum_1^p \omega_i = 1$, be fixed normalized weights, which determine the weighted means statistic, $\hat{\mu} = \sum_1^p \omega_i Y_i$, with the variance, $Var(\hat{\mu}) = \sum_1^p \omega_i^2 Var(Y_i)$. Then

$$Var\,(Y_k - \hat{\mu}) = (1 - 2\omega_k)Var(Y_k) + \sum_1^p \omega_i^2 Var(Y_i). \tag{4.9}$$

In the case when $\omega_i = \omega_i^0 = w_i^0 \left[\sum_{j=1}^p w_j^0\right]^{-1}$, $\hat{\mu}$ is the optimal least squares estimator, and the second term in the right-hand side of (4.9) simplifies to $\left[\sum_{i=1}^p Var(Y_i)^{-1}\right]^{-1} = \omega_k^0 Var(Y_k)$. By substituting this expression in (4.9), one obtains $Var\,(Y_k - \hat{\mu}) = (1-\omega_k^0)Var(Y_k)$. Horn, Horn and Duncan (1975, p 382) argue that by continuity, if the weights are close to ω_k^0, this is an approximate identity. Thus, one derives an *almost unbiased* estimator of $Var(Y_k)$ as $(Y_k - \hat{\mu})^2/(1-\omega_k)$, and the corresponding estimate of the variance, $Var(\hat{\mu})$, is $\sum_1^p \omega_i^2 (Y_i - \hat{\mu})^2/(1-\omega_i)$.

This statistic gives an estimate of the variance of any weighted means statistic for the weights (4.4) when s_i^2 are fixed. The method for the plug-in weights $\omega_i = (z + \sigma_{B_i}^2 + s_i^2)^{-1}/\sum_k (z + \sigma_{B_k}^2 + s_k^2)^{-1}$ leads to the following estimate of the variance of $Var(\hat{\mu})$,

$$\widehat{Var}(\hat{\mu}) = \left[\sum_{j=1}^p \frac{1}{z + \sigma_{B_j}^2 + s_j^2}\right]^{-1} \sum_{i=1}^p \frac{(Y_i - \hat{\mu})^2}{(z + \sigma_{B_i}^2 + s_i^2)^2} \left[\sum_{k:k\neq i} \frac{1}{z + \sigma_{B_k}^2 + s_k^2}\right]^{-1}. \tag{4.10}$$

Simulations, some of which are reported in the next Section, show that (4.10) provides a good approximation to the true value of this variance for random weights above with $z = z_{DL}$, $z = z_{MP}$, or $z = z_W$. They demonstrate that (4.10) is superior to the estimate,

$$\left[\sum_1^p \frac{1}{z + \sigma_{B_i}^2 + s_i^2}\right]^{-1}, \tag{4.11}$$

which has been suggested by Mandel (1991, p 72), $z = z_{MP}$, by DerSimonian and Laird (1986, p 183), $z = z_{DL}$, or by Willink (2002, p 348), $z = z_W$. Moreover, it gives a better estimate of the variance of the maximum likelihood estimator than the inverse of the observed Fisher information $[\sum_1^p 1/(z + \sigma_{B_i}^2 + \hat{\sigma}_i^2)]^{-1}$ (which typically underestimates the true variance.) For large p, it is close to the estimate suggested by Rukhin and Vangel (1998),

$$\sum_1^p (Y_i - \hat{\mu})^2 (z + \sigma_{B_i}^2 + s_i^2)^{-2} [\sum_{k=1}^p 1/(z + \sigma_{B_k}^2 + s_k^2)]^{-2}.$$

Simulations also demonstrate that the tails of the distribution of the pivotal ratio, $(\hat{\mu} - \mu)/\sqrt{\widehat{Var}(\hat{\mu})}$ can be well approximated by those of the t-distribution with $p-1$ degrees of freedom.

The same method produces uncertainty estimates of the square of the kth laboratory deviation from the KCRV. This deviation, the so-called degree of equivalence, must be reported according to MRA (1999), which does not give a formal definition. For a given KCRV estimator $\hat{\mu}$ we define the degree of equivalence of the kth laboratory as $E(Y_k - \hat{\mu})$. In the model (3.1), when $\hat{\mu}$ is a weighted means statistic, $E(Y_k - \hat{\mu}) = 0$. Thus, the statistic $(Y_k - \hat{\mu})^2, k = 1, \ldots, p$, captures possible violations of (3.1). By using (4.9), the almost unbiased estimate of $Var(Y_k)$, as above, produces the following estimator,

$$\widehat{E}(Y_k - \hat{\mu})^2 = \frac{(1 - 2\omega_k)(Y_k - \hat{\mu})^2}{1 - \omega_k} + \sum_1^p \frac{\omega_i^2 (Y_i - \hat{\mu})^2}{1 - \omega_i}$$

$$= (1 - \omega_k)(Y_k - \hat{\mu})^2 + \sum_{i \neq k} \frac{\omega_i^2 (Y_i - \hat{\mu})^2}{1 - \omega_i}.$$

Notice that $\widehat{E}(Y_k - \hat{\mu})^2 > (1 - \omega_k)(Y_k - \hat{\mu})^2$. Thus the ratio, $T_k = (1 - \omega_k)$ $(Y_k - \hat{\mu})^2 / \widehat{E}(Y_k - \hat{\mu})^2$, is bounded, $0 < T_k < 1$, and, according to our simulations, has approximate beta-distribution whose parameters can be estimated by the method of moments. The values of T_k exceeding the critical point of this beta-distribution suggest that EY_k is different from μ.

In Section 5 results of a Monte Carlo simulation study are reported for several weighted means statistics; the variances of these statistics were estimated via (4.10). In addition, the maximum likelihood estimate of μ and the median, which do not have the form (4.1) with weights (4.4), are also evaluated there. To estimate the variance of the median, the method of Sheather (1986) has been used. The corresponding estimator is based on order statistics, $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(p)}$,

$$\sum_1^p w_k (Y_{(k)} - \tilde{Y})^2, \tag{4.12}$$

with $\tilde{Y} = \sum_k w_k Y_{(k)}$ and probabilities w_k given in Tables 1 and 3 of Sheather (1986). For example, when p is odd, $(p = 2m + 1)$

$$w_k = \frac{\left(\frac{k-1/2}{p}\right)^m \left(\frac{p-k+1/2}{p}\right)^m}{\sum_{j=1}^p \left(\frac{j-1/2}{p}\right)^m \left(\frac{p-j+1/2}{p}\right)^m}.$$

This method gives more accurate variance estimates than that of Maritz and Jarrett (1978), but still produced upwardly biased estimates. This fact is not surprising in view of this method's close relationship to the jackknife procedure, which can be inconsistent in our situation (Efron and Tibshirani, 1993, Sec 11.6). The variance of the maximum likelihood estimator was estimated as the inverse of the observed Fisher information; this statistic underestimated the true variance.

5. Example and Simulation Results

Here the results of a Monte Carlo simulation study when $p = 11$; $\mu = 0, \sigma_L^2 = 0, 2, 4$; and $\eta^2 = 0 : 0.25 : 4$ are reported. The variance component, $\sigma_{B_i}^2$, of the type B uncertainty was taken to be proportional to a χ^2-random variable with $\nu = 1, 4$ or 12 degrees of freedom, $\sigma_{B_i}^2 \sim q\chi^2(\nu), E\sigma_{B_i}^2 = q\nu$. We took the distribution of the within laboratory variances, σ_i^2, to be the inverted gamma-distribution $1/\Gamma(\alpha, \beta)$ with parameters $\alpha = 2$, $\beta = 1$, so that $E\sigma_i^2 = 1$. The sample means Y_i were simulated as $Y_i = \eta + \sqrt{\sigma_L^2 + \sigma_i^2 + \sigma_{B_i}^2} Z_i$ from a standard normal sample $Z_1, \ldots, Z_p$; the sample variances s_i^2 were taken to be realizations of multiples of χ^2-random variables, $s_i^2 \sim \sigma_i^2 \chi^2(n_i - 1)/(n_i - 1)$.

We studied the mean squared errors (mse) of the unweighted mean of the sample means, $\bar{Y}$; the median, $med(Y)$; the Graybill-Deal estimator (**??**); the version of Graybill-Deal estimator, GD_0 which does not include $\sigma_{B_i}^2$, but is based on the weights proportional to s_i^{-2}; the DerSimonian-Laird estimator (4.6); the Mandel-Paule estimator (4.7); the version of the maximum likelihood estimator (4.8); and the maximum likelihood estimator, ML which was implemented by a modification of the iterative algorithm involving solutions of cubic equations for σ_i^2 as described in Vangel and Rukhin (1999).

Table 1: The mean squared errors (mse) when $n = 3, 10, \sigma_L^2 = 0, 2, 4$, and $\eta^2 = 0$

	$n = 3$			$n = 10$		
	$\sigma_L^2 = 0$	$\sigma_L^2 = 2$	$\sigma_L^2 = 4$	$\sigma_L^2 = 0$	$\sigma_L^2 = 2$	$\sigma_L^2 = 4$
$\bar{Y}$	1.18	1.37	1.56	1.11	1.30	1.48
med(Y)	1.35	1.68	2.03	1.33	1.71	2.02
GD	0.79	1.09	1.39	0.76	1.07	1.37
GD_0	4.29	4.92	5.62	1.64	2.29	2.60
DL	0.81	1.06	1.31	0.79	1.03	1.23
MP	0.81	1.07	1.31	0.79	1.04	1.24
W	0.76	1.06	1.26	0.73	0.88	1.17
ML	0.72	1.02	1.29	0.70	0.81	1.12

The results are collected in Table 1 when $n_i \equiv n = 3, 10$; only results for $q = 3, \nu = 4$ are reported there. The increase in ν for a fixed product $q\nu$ leads to smaller mse, but also to smaller coverage probabilities. Table 1 contains the mean squared errors of the studied eight estimators. Since all estimators have the bias η, mse is the sum of η^2 and the variance (which is the mean squared error when $\eta = 0$.) For this reason Table 1 reports the results only when $\eta = 0$. Neither the median nor the version GD_0 of the Graybill-Deal estimator were ever the best in terms of the mean squared error, while the Graybill-Deal estimator, which includes Type B uncertainty, the DerSimonian-Laird method, the Mandel-Paule algorithm (which behaved very similarly to the DerSimonian-Laird method) systematically were among the best procedures. They did remarkably well compared to the "golden standard" of the maximum likelihood procedure which is much more computationally intensive. For small values of n these estimators are very close to the maximum likelihood estimator, whose behavior is similar to that of the Willink version of the maximum likelihood procedure (4.8).

Table 2: The mean squared errors (mse) for random $n, \sigma_L^2 = 0, 2, 4, \eta^2 = 0$ and their estimates via (4.10) and (4.11)

	$\nu = 3, q = 4$			$\nu = 1, q = 12$		
	$\sigma_L^2 = 0$	$\sigma_L^2 = 2$	$\sigma_L^2 = 4$	$\sigma_L^2 = 0$	$\sigma_L^2 = 2$	$\sigma_L^2 = 4$
$\bar{Y}$	1.18	1.35	1.54	1.18	1.36	1.54
(4.10)	1.18	1.35	1.54	1.17	1.36	1.54
med(Y)	1.25	1.62	1.97	0.68	1.15	1.54
(4.12)	1.65	2.09	2.46	1.13	1.58	2.02
GD	0.73	1.03	1.38	0.33	0.87	1.21
(4.10)	0.70	0.98	1.26	0.29	0.68	1.07
(4.11)	0.65	0.70	0.70	0.26	0.27	0.28
DL	0.76	1.00	1.24	0.32	0.68	0.93
(4.10)	0.75	0.98	1.20	0.32	0.61	0.85
(4.11)	0.85	1.02	1.21	0.38	0.61	0.85
MP	0.75	1.00	1.25	0.33	0.73	1.02
(4.10)	0.75	1.00	1.26	0.33	0.69	1.01
(4.11)	0.80	1.05	1.21	0.29	0.62	0.84
W	0.73	1.02	1.30	0.31	0.83	1.22
(4.10)	0.73	0.97	1.22	0.29	0.67	0.96
(4.11)	0.72	0.80	0.95	0.27	0.31	0.42
ML	0.73	0.96	1.19	0.31	0.64	0.89
(4.10)	0.70	0.94	1.19	0.27	0.61	0.88
(4.11)	0.70	0.95	1.10	0.26	0.61	0.88

These results also hold in the studied unbalanced cases. Table 2 gives the simulated values of mse when the sample sizes of $p = 11$ laboratories are permutations of integers from 3 to 13, $q = 3, \nu = 4$, in the first three columns, $q = 1, \nu = 12$, in the last three columns, $\sigma_L^2 = 0, 2, 4, \eta^2 = 0$. Their estimates via (4.10) and (4.11), as discussed in Section 4, are also provided. Clearly (4.10) gives a much better estimate of the mse of the weighted means procedures than (4.11). For the reason indicated above, the estimator GD_0 was omitted in this Table as well as in Figures 1 and 2 which depict characteristics of only five estimators.

Figure 1 displays the confidence coefficient of the intervals $\hat{\mu} \pm 2\sqrt{\widehat{Var}(\hat{\mu})}$, for these estimators when $\eta^2 = 0 : 0.25 : 4$. The confidence coefficient of the interval obtained from all estimators, except the median, fell below 0.60 for $\eta^2 = 4$.

The average half-widths (standard errors) of exact 95%-confidence intervals are depicted in Figure 2. They are increasing as η increases, but not fast enough to compensate for the loss in confidence.

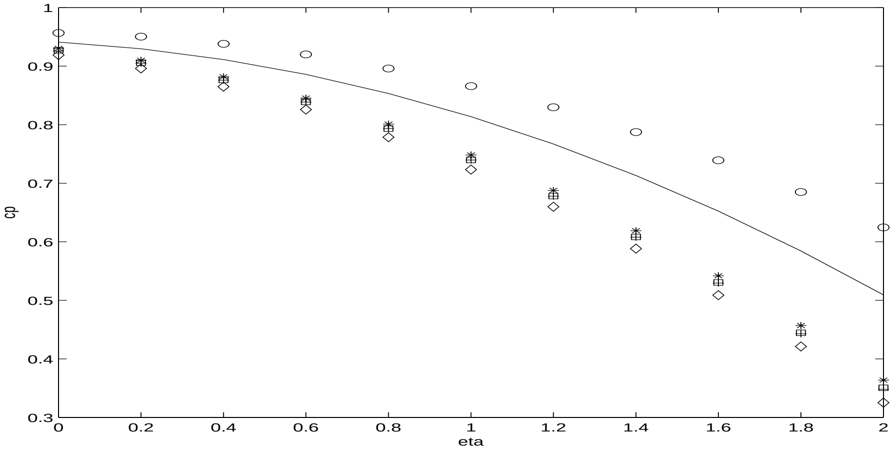

Figure 1: Plot of coverage probability of confidence intervals $\hat{\mu} \pm 2\sqrt{\widehat{Var}(\hat{\mu})}$ vs η for five estimators (the continuous line corresponds to $\bar{Y}$, the line marked by 'o' to $med(Y)$, '*' line to DL , '+' line to W, the diamonds line to ML.)

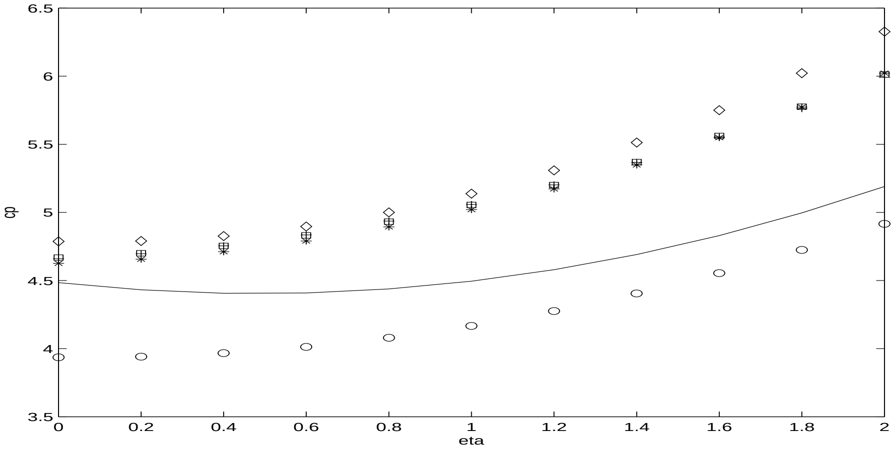

Figure 2: Plot of standard errors of confidence intervals $\hat{\mu} \pm 2\sqrt{\widehat{Var}(\hat{\mu})}$ vs η for five estimators (designations of lines are the same as in Figure 1).

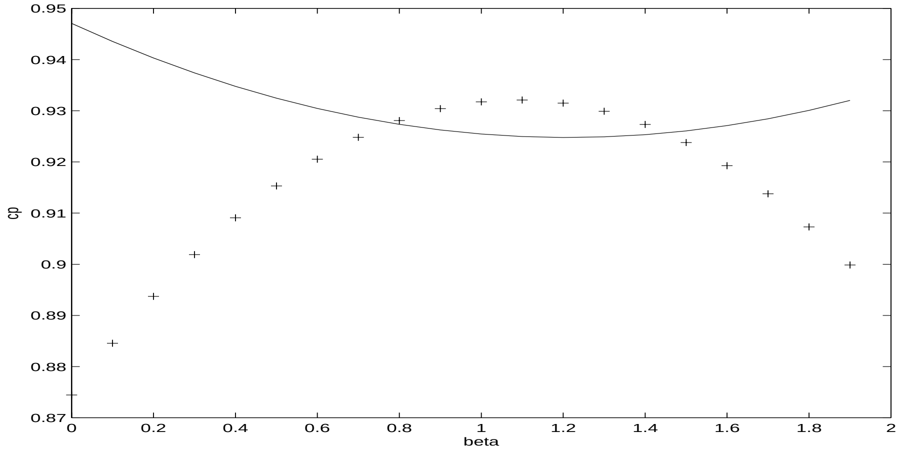

Figure 3: Plot of coverage probability of confidence intervals $\hat{\mu} \pm 2\sqrt{\widehat{Var}(\hat{\mu})}$ vs β for the DerSimonian-Laird estimator (the continuous line corresponds to $\bar{Y}$)

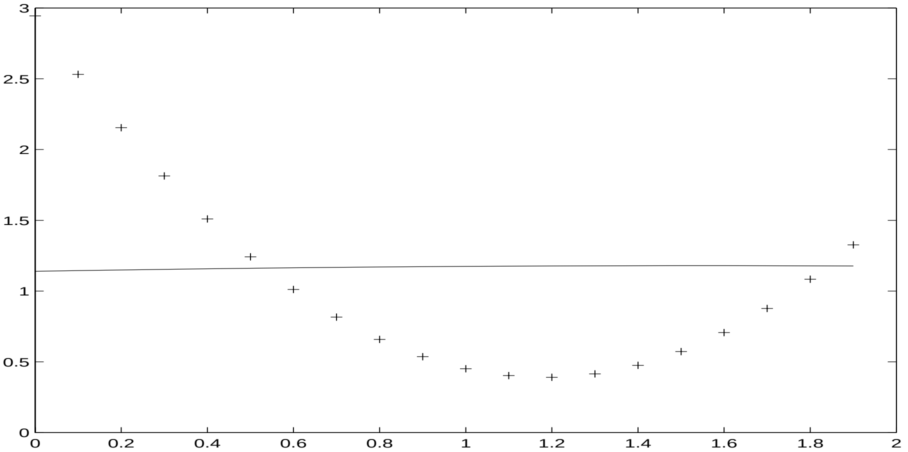

Figure 4: Plot of mean squared errors vs β for the DerSimonian-Laird estimator and $\bar{Y}$

These facts confirm the deteriorating performance of the confidence intervals for large bias component. Clearly, all characteristics are very influenced by this parameter.

To investigate robustness of these procedures with regard to misspecification of Type B error we evaluated the mean squared error and the confidence coefficient of the interval $\hat{\mu} \pm 2\sqrt{\widehat{Var}(\hat{\mu})}$, when one of the laboratories reports a multiple $\beta\sigma^2_{B_i}$, $\beta = 0 : 0.2 : 2$, instead of the true type B uncertainty $\sigma^2_{B_i}$. The corresponding results for the DerSimonian-Laird estimator (4.6) contrasted by $\bar{Y}$ are displayed in Figures 3 and 4 when $\eta = 0$. Both of these characteristics are seriously affected by under-reporting of the type B uncertainty. Figure 5 shows the empirical distribution of the pivotal ratio, $(\hat{\mu}-\mu)/\sqrt{\widehat{Var}(\hat{\mu})}$, and its approximation by a t-distribution with $p-1$ degrees of freedom when $\eta=0$.

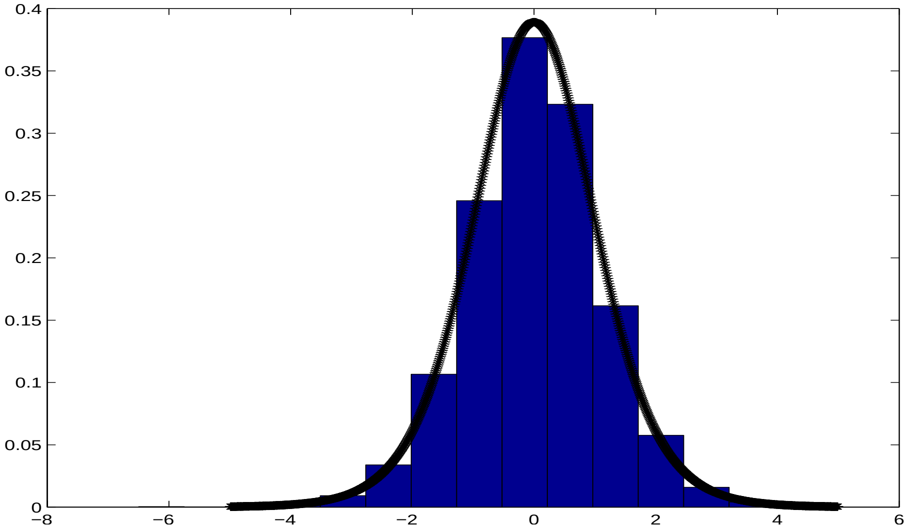

Figure 5: The histogram of the distribution of the pivotal ratio, $(\hat{\mu}-\mu)/\sqrt{\delta_0}$ and its t-density approximation.

For illustration, these techniques were implemented in the accelerometers key comparisons study (CCAUV.V-K1) described in Section 2. Only the results for the frequencies $40, 50$ and 63Hz and single-ended accelerometers are given here. Table 3 contains approximate confidence intervals obtained by techniques of Section 4 applied to CCAUV.V-K1. As in the simulation study, the median and the version GD_0 of the Graybill-Deal estimator provide the least satisfactory answers. The Graybill-Deal estimator, which includes Type B uncertainty, the Mandel-Paule procedure and the DerSimonian-Laird method all give the same answer which practically coincides with the PTB solution reported by von Martens *et el.* (2002).

Table 3: The confidence intervals $\hat{\mu}\pm 2\sqrt{\delta_0}$ for eight estimators in CCAUV.V-K1 study, frequencies $40, 50, 63$ Hz

$\bar{Y}$	med(Y)	GD	GD$_0$	DL	MP	W	ML
0.12894	0.12884	0.12998	0.12886	0.12898	0.12898	0.12898	0.12898
± 0.00022	± 0.00034	± 0.00012	± 0.00019	± 0.00012	± 0.00012	± 0.00013	± 0.00011
0.12896	0.12896	0.12999	0.12885	0.12899	0.12899	0.12895	0.12899
± 0.00016	± 0.00054	± 0.00010	± 0.00014	± 0.00010	± 0.00010	± 0.00013	± 0.00010
0.12890	0.12886	0.12896	0.12886	0.12896	0.12896	0.12895	0.12896
± 0.00018	± 0.00028	± 0.00010	± 0.00016	± 0.00010	± 0.00010	± 0.00010	± 0.00011

6. Conclusions

Under the model (3.1) the considered estimators provide, on average, similar values for KCRV but with differing uncertainties. However, while weighted means estimators profit from dependence on Type B uncertainty in terms of stability when the epistemic errors are reported correctly, this dependence also compromises the performance of these estimators when the Type B uncertainty is misstated or when the variability inherent in the stated error is ignored.

Two particular circumstances could lend themselves to misstatement of Type B uncertainty: (i) incorporation of "offsets" into NMI's reported measurements and (ii) differing evaluation of epistemic uncertainty by different scientists, even under identical conditions, using the same equipment and procedures. "Offsets" are standard adjustments (usually additive in the reporting scale for the measurements) based on reproducible differences that are historically present between NMIs without other apparent explanation. Scientists acknowledge these systematic differences by defining an offset to effectively "reset the scales to zero" for the NMIs involved. Many scientists also recognize that an offset is known only to a certain precision (i.e. has an associated uncertainty of its own.) The incorporation of offsets partially corrects for a bias that would otherwise be present, but does so at the expense of additional uncertainty and/or residual bias.

The differences in the uncertainty budgets predominantly reflect epistemic uncertainty discrepancies. By necessity, each NMI provides *its own* expert judgment about *its own* measurements, i.e., the NMI's own warrant of the nature of its measurements. So the stated Type B uncertainty for each participant reflects both the epistemic uncertainty for that participant's measurements and also individuality of its expert opinion. Even for an identical measurement methodology, experts can vary widely in their assessments although all uncertainty budgets for inter-comparison encompass the same list of specified factors. The differences in personal opinion about specific components or about their aggregation to yield Type B uncertainty can result in wide dissimilarities. Thus, when the method for calculating the KCRV depends heavily on epistemic uncertainty, NMIs with apparently similar processes may contribute to the KCRV differently by virtue of their diverse weights. A direct consequence of the failure to include in the KCRV uncertainty calculations any representation of either the uncertainty of a reported systematic effect (e.g. offset) could lead to substantial bias in the KCRV. While the effect of the variation in a reported type B uncertainty is more benevolent, still it is the potential misrepresentation of KCRV in the direction of values reported by NMIs with the smallest stated epistemic uncertainty. The Graybill-Deal estimator, especially its version GD_0, which does not include $\sigma^2_{B_i}$, is particularly prone to such a misrepresentation.

In the "non-regular" case, where NMIs are actually divergent and model (3.1) is not applicable, this consequence could be magnified, depending upon the divergence of the NMIs and their relative values for (non-normal) type B uncertainty. It is easy to conjecture that departures from Gaussian to asymmetric distributions, particularly with heavy tails, could further accentuate the influence of NMIs stating smallest Type B uncertainties and yielding substantially understated uncertainty for the KCRV. Consequently, a *rigorous statistical approach* is even more important in the non-regular case.

Appendix

It suffices to prove (4.2) when $\sum_1^p w_i = 1$. Then

$$E\sum_1^p w_i(Y_i-\hat{\mu})^2 = \sum_1^p w_i\left[(1-w_i)^2 Var(Y_i) + \sum_{k\neq i} w_k^2 Var(Y_k)\right]$$

$$= \sum_1^p w_i\left[(1-2w_i)Var(Y_i) + \sum_k w_k^2 Var(Y_k)\right]$$

$$= \sum_1^p w_i Var(Y_i) - \sum_1^p w_i^2 Var(Y_i),$$

which is the desired identity.

References

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments, *Journ. Royal Statist. Soc.*, Supplement **4**, 102-118.

Cochran, W. G. (1954). The combination of estimates from different experiments, *Biometrics*, 10 ,101-129.

Comite International des Poids et Mesures. (1999). Mutual recognition of national measurement standards and measurement certificates issued by National Metrology Institutes, Technical Report, Sevres, France.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman&Hall, New York.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials, *Controlled Clinical Trials* **7**, 177-188.

Gleser, L. J. (1998). Assessing uncertainty in measurement, *Statist. Sci.* **13**, 277-290.

Hall, B. D. and Willink, R. (2001). Does "Welch-Satterthwaite" make a good uncertainty estimate? *Metrologia* **38**, 9-15.

Harville, D. (1977). Maximum likelihood approaches to variance components estimation and related problems, *Journ. Amer. Statist. Assoc.* **72**, 320-340.

Horn, R. A., Horn, S. A. and Duncan, D. B. (1975). Estimating heteroscedastic variance in linear models, *Journ. Amer. Statist. Assoc.* **70**, 380-385.

Mandel, J. (1991). *Evaluation and Control of Measurements*, M. Dekker.

Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median, *Journ. Amer. Statist. Assoc.* **73**, 194-196.

Martens, von H. J. *et el.* (2002). Report on key comparison CCAUV.V-K1, *PTB-1.22*. Braunschweig, Germany.

Mohr, P. J. and Taylor, B. N. (1999). CODATA recommended values of the fundamental physical constants, *Journ. Physical Chem. Ref. Data* **28**, 1713-1852.

Paule, R. C. and Mandel, J. (1982). Consensus values and weighting factors, *Journ. Res. Nat. Inst. Stand. Techn.* **87**, 377-385.

Rao, P. S. R. S. (1981). Cochran's contributions to variance component models for combining estimates, In *W.G. Cochran's Impact on Statistics* (Edited by P. Rao and J. Sedransk, editors,). J. Wiley.

Rukhin, A. L. (2003). Two procedures of meta-analysis in clinical trials and interlaboratory studies, *Tatra Mount. Math. Publ.* **28**, 155-168.

Rukhin, A. L., Biggerstaff, B. and Vangel, M.G. (2000). Restricted maximum likelihood estimation of a common mean and Mandel-Paule algorithm, *Journ. Statist. Plan. Inf.* **83**, 319-330.

Rukhin, A. L. and Vangel, M.G. (1998). Estimation of a common mean and weighted means statistics, *Journ. Amer. Statist. Assoc.* **93**, 303-308.

Searle, S., Casella, G. and McCulloch, C. (1992). *Variance Components*. J. Wiley.

Sheather, S. J. (1986). A finite sample estimate of the variance of the sample median, *Statist.&Probab. Lett.* **4**, 337-342.

Swyt, D. (2001). Length and dimensional measurements at NIST, *Journ. Res. Nat. Inst. Stand. Techn.* **106**, 1-23.

Vangel, M. G. and Rukhin, A. L. (1999). Maximum likelihood analysis for a series of similar experiments, *Biometrics* **55**, 302-313.

Willink, R. (2002). Statistical determination of a comparison reference value using hidden errors, *Metrologia* **39**, 343-354.

Received January 22, 2006; accepted May 22, 2006.

Andrew L. Rukhin
Statistical Engineering Division
National Institute of Standards and Technology
Building 820
Gaithersburg MD 20899 USA
rukhin@cam.nist.gov
rukhin@math.umbc.edu

N. Sedransk
Department of Mathematics and Statistics
University of Maryland at Baltimore County
1000 Hilltop Circle
Baltimore MD, 21250 USA
sedransk@niss.org