

HOW TO COMBINE RESULTS HAVING STATED UNCERTAINTIES: TO MU OR NOT TO MU?

Published in: Ales Fajgelj, Maria Belli, Umberto Sansone (Eds). *Combining and reporting analytical results*. Royal Society of Chemistry, London (2007), pp 127-142.

David L. Duewer

Analytical Chemistry Division, Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8390 USA

1 INTRODUCTION

It is widely accepted that knowledge of the expected uncertainty of a measured value is necessary for judging the measurement's fitness-for-purpose.¹ Considerable effort continues to be expended on how to best estimate and report measurement uncertainty (MU) for chemical and biological measurands.²⁻⁴ Several academic research groups are actively exploring ways of using MU estimates in multivariate data analysis.^{5,6} It is therefore perhaps surprising that some experienced data analysts argue against using MU in even relatively simple univariate applications such as using interlaboratory data to assign the expected value of a measurand in particular materials.⁷⁻⁹ But if MU is not expected to be useful for univariate tasks, can it be useful in multivariate applications? More generally, can estimates of MU provide quantitatively useful chemical information when there may be significant bias among the data analyzed, such as when measurements are from more than one source or are taken over a long period of time?

The primary concern of those who advise against using MU in establishing consensus values is that estimating MU is not yet sufficiently routine and standardized to ensure that different participants interpret and report the uncertainties in their values in a consistent manner. Any enquiry into the potential utility of MU for assigning consensus values to interlaboratory study materials therefore requires, at a minimum, that there be philosophically consistent MU estimates for all of the reported measurements and, to enable valid comparison, trustworthy reference values (RVs) for the studied measurands. Few, if any, "routine" interlaboratory studies currently meet this standard. Indeed, reliably estimating even short-term measurement precision from studies involving well-motivated but metrologically naïve participants is challenging.¹⁰

Beginning in the 1993, a number of very non-routine interlaboratory studies involving various chemical measurands have been performed under the auspices of the Comité Consultatif pour la Quantité de Matière (CCQM) of the Comité International des Poids et Mesures (CIPM). The number and diversity of these studies dramatically increased with the signing in late 1999 of the Mutual Recognition Arrangement (CIPM MRA) for national measurement standards and for calibration and measurement certificates issued by national metrology institutes (NMIs). Currently, data are publicly available for 150 measurands evaluated in 33 studies performed by the member organizations of the

CCQM. These organizations have considerable expertise in chemical metrology and experience with MU evaluation. All of the samples used in the CCQM studies are thoroughly characterized for homogeneity and stability. Virtually all of the individual values are reported with a symmetrical uncertainty estimate of defined confidence. The source of every datum in these studies is fully attributed. Most importantly, reliable RVs – assigned either by gravimetric sample preparation or extensively debated expert review – are available for all measurands.

This report describes the use of these publicly available CCQM data to explore whether appropriately evaluated MU estimates can be useful in combining results from interlaboratory studies. Three different approaches to using MU are evaluated (weighting, bootstrap resampling, and mixture-models) relative to the performance of a number of commonly used or recently proposed evaluation metrics.

2 DATA

2.1 CCQM Key Comparisons and Pilot Studies

As of February 2006, data are publicly available from the 23 CCQM Key Comparisons (KCs) and 10 CCQM pilot studies listed in Table 1.^{11,12} KCs are formal studies designed to establish the extent of measurement agreement among the participating CCQM member organizations and their official delegates for particular measurands. All of the measurement values submitted for a particular KC are made public with the formal acceptance of the study's Final Report. Depending on the type of sample and measurands investigated, some form of RV is established for all measurands. Many of these RVs are assigned through gravimetric preparation verified by measurement, others are assigned through a combination of consensus technical judgment and statistical analysis. No participant's KC result can be withdrawn from publication once it has been submitted; however, results that are technically suspect are generally not used in assigning RVs.

Pilot studies are relatively less formal investigations that may also involve expert but non-CCQM member participants. Pilot studies often are designed to investigate technical challenges in particular measurement processes or evaluate the relative metrological utility of alternate measurement technologies. They are sometimes conducted in parallel with KCs, using the same sample materials, timing, and shipping protocols, to allow interested parties to evaluate their measurement capabilities without the burden of public participation. While all pilot study results are available to all of the participants in the study, every participant must agree to publication for attributed study results to be made public. In addition to such separate publication, data for a few pilot studies have been published in the Final Report of a follow-on KC addressing similar measurands. While pilot studies are typically not as thoroughly catechized as are KCs, well-evaluated RVs are available for all of the measurands in the published pilot studies.

Table 1 lists the designation code ("Study"), a general description of the measurands ("Measurand(s)"), the number of participants in each study ("#Labs"), the total number of different measurands reported in the study ("#Sets"), and whether the RVs are assigned through gravimetric preparation of the sample materials. The number of participants is listed as a range for some studies that address multiple analytes, multiple samples, and/or different measurement conditions. In all, a total of 1294 individual measurement values are available.

Table 1 *CCQM Key Comparison and Pilot Studies Publicly Available as of 1-Feb-2006*

Study	Measurand(s)	#Labs	#Sets	Grav?
K01	Trace gases in N ₂ , components of natural gas	8 - 10	28	All
K02	Metals in natural water	9	2	
K03	Gases in auto emissions	13	3	All
K04	Ethanol in air	8	1	All
K05	pp'-DDE in fish oil	10	2	
K06	Cholesterol in serum	7	2	
K07	Volatile organic compounds in N ₂	8	5	All
K08	Metal calibration solutions	12 - 13	4	All
K09	pH of phosphate buffer	4 - 10	20	
K10	Volatile organic compounds in N ₂	8	3	All
K11	Total glucose in human serum	3	1	
K12	Creatinine in human serum	5	2	
K13	Metals in sediment	14	2	
K14	Calcium in serum	9	1	
K16	Components of natural gas	7 - 9	23	All
K17	pH of phthalate buffer	11	3	
K21	pp'-DDT in fish oil	8 - 9	2	
K24	Cadmium in rice	18	1	
K25	PCBs in sediment	7 - 9	5	
K27	Ethanol in aqueous matrix	2 - 9	7	6 of 7
K28	Tributyltin in sediment	10	2	
K31	Arsenic in shellfish	7	1	
K43	Metals in salmon	5 - 9	5	
P06	Cholesterol in serum	7	2	
P08	Glucose in serum	4	2	
P09	Creatinine in serum	5	2	
P13	Metals in artificial food digest	10 - 12	3	All
P17	PCBs in sediment	8 - 10	4	
P18	Tributyltin in sediment	10 - 12	2	
P29	Zinc in rice	14	1	
P32	Anion calibration solutions	9 - 11	2	All
P39	Metal in tuna fish	8 - 15	5	
P43	Dibutyltin in sediment	11	2	

2.2 Participants

Only signatories of the CIPM MRA and their official designees can participate in KC studies. Therefore, more than 95 % (1230 of the 1294 total) of the publicly available CCQM data are from the 31 national or international organizations listed in Table 2. The remaining 62 pilot study measurements are reported by 17 different academic, commercial, or governmental laboratories that have interest and expertise in the particular measurement systems studied.

2.3 Measurements

Measurement results for all KCs must be reported as values with an associated MU of

Table 2 *NMI Participation in the Public CCQM Key Comparison and Pilot Studies*

Organization(s)	Representing	#Values
BAM, PTB, UBA	Germany	144
NIST	USA	122
VNIIM, VNIIFTRI	Russia	114
NMIJ	Japan	106
KRISS	South Korea	104
LGC, NPL	United Kingdom	96
NRCCRM	PR China	84
NMi-VSL	The Netherlands	76
LNE	France	70
NMIA	Australia	58
GUM	Poland	46
SMU	Slovak Republic	40
OMH	Hungary	35
IRMM	European Union	30
NRC	Canada	30
CENAM	Mexico	27
DPL, Radiometer	Denmark	12
EMPA, METAS	Switzerland	10
CSIR-NML	South Africa	8
IAEA	Global	5
IMGC	Italy	5
NPLI	India	4
CMI	Czech Republic	3
INMETRO	Brazil	2
SP	Sweden	1
		1232

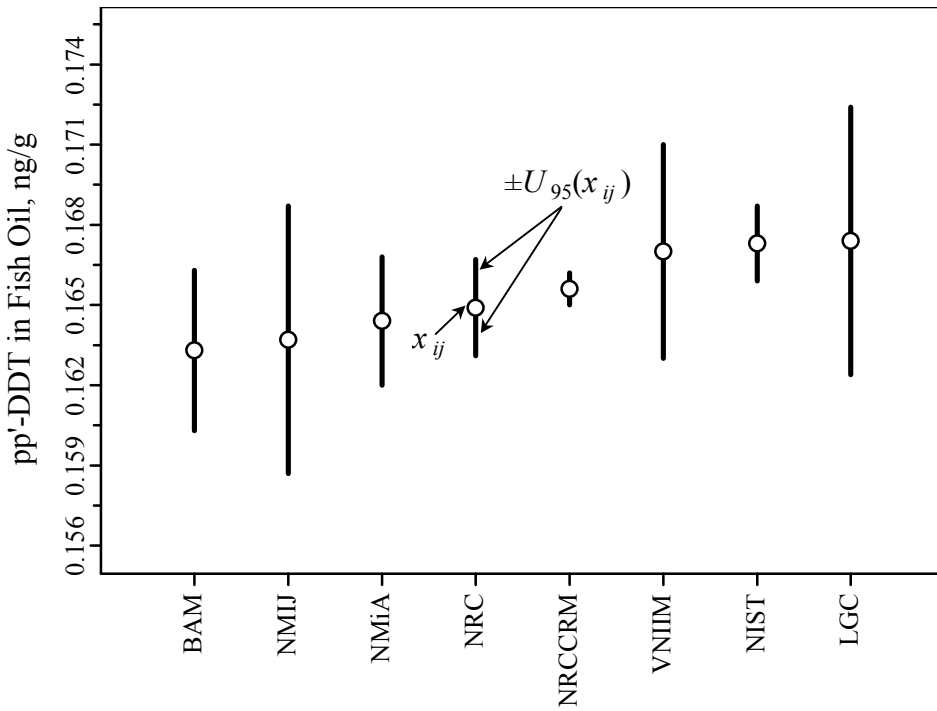
defined confidence. For all of the data in the published CCQM studies, the confidence intervals are symmetric about the expected value and are expanded with the intent to include the true value with about 95 % confidence. The data for the i^{th} participant in the j^{th} data set can thus be denoted: $x_{ij} \pm U_{95}(x_{ij})$. These data are typically displayed using standard “dot-and-bar” graphs such as that shown in Figure 1.

Since all of the MUs in these CCQM studies are expressed as symmetric 95 % confidence intervals, the CCQM measurements can also be interpreted as normal kernel densities with expected value x_{ij} and standard deviation proportional to $U_{95}(x_{ij})$ but covering the true value with about 68 % confidence, $U_{68}(x_{ij})$. For normal distributions

$$U_{68}(x_{ij}) \cong U_{95}(x_{ij})/2 . \quad (1)$$

Just as each measurement kernel, $N(x_{ij}, U_{68}(x_{ij}))$, represents the expected probability density function (PDF) of the true location given the measurement, the sum of the n_j kernels defines a mixture-model PDF (MM-PDF) for the combined set of measurements⁹

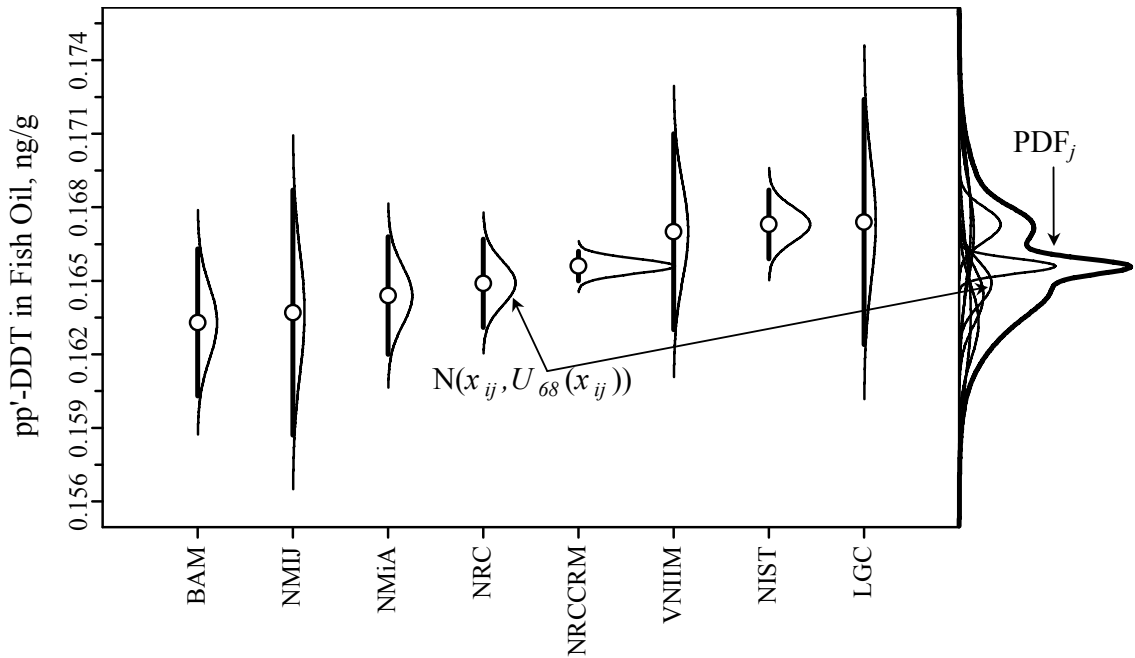
$$\text{PDF}_j = \sum_i^{n_j} N(x_{ij}, U_{68}(x_{ij})) / n_j . \quad (2)$$



Participants in CCQM-K21, Sample B

Figure 1 Measurements with Stated Uncertainties as Dot-and-Bars, $x_{ij} \pm U_{95}(x_{ij})$

This alternate kernel density and PDF representation of the CCQM measurements is displayed in Figure 2. The PDF_j at the right of the graph is a marginal distribution for the combined results.



Participants in CCQM-K21, Sample B

Figure 2 Measurements with Stated Uncertainties as Kernel Densities, $N(x_{ij}, U_{68}(x_{ij}))$

3 LOCATION METRICS

There are a very large number of summary statistics that are routinely used or have been proposed for estimating a measurand's "true value" from interlaboratory study results. Each of these location metrics makes somewhat different assumptions about the nature of the data being combined and summarized. The extent of agreement between the estimated location and "truth" depends upon how well the observed data match these underlying assumptions. Only a few, hopefully representative, metrics are examined in this report. Graphical representations of the philosophical basis for most of these metrics are presented elsewhere.¹³

Location metrics are often examined with respect to their *efficiency* (how rapidly the estimate converges upon the true value as the number of observations increases, assuming that all data accord with the metrics assumptions) and their *robustness* (how insensitive the metrics are to violations of the assumptions).¹⁴ While there are a number of ways to quantitatively evaluate these properties for specific situations, in this report these two properties are only qualitatively discussed.

3.1 "Traditional" Metrics That Do Not Utilize Measurement Uncertainties

3.1.1 Mean. Used in conjunction with some form of "outlier" rejection, the mean or arithmetic average is the most commonly used location metric. The mean is readily calculated in the familiar closed form

$$\text{Mean}_j = \sum_i^{n_j} x_{ij} / n_j . \quad (3)$$

where n_j is the number of results to be combined and summarized. When all of the results are truly random draws from a population that is normally distributed with location μ and standard deviation σ , denoted $N(\mu, \sigma)$, then the mean is the most efficient of the location estimates. The expected difference between the mean of the sample and μ declines as the square-root of the number of observations increases. The mean is fairly insensitive to the detailed shape of distribution as long as the observed results are symmetrically distributed about μ . However, the mean is not robust to the presence of unrecognized "outlier" results or asymmetric sampling of the underlying population. Recognizing this sensitivity, the mean is generally used to summarize results after some form of "outlier" data rejection has been used to identify and remove suspect data.

3.1.2 Median. The median is the value that is in the center of a given set of results. It is the most common "robust" location metric and has been recommended for use with KC data.^{15,16} The median is readily calculated from rank-ordered data

$$\begin{aligned} \text{Median}_j &= x_{kj}, k = \text{int}(n_j/2) + 1 \text{ when } n_j \text{ is odd} \\ &= (x_{kj} + x_{k+1j})/2, k = \text{int}(n_j/2) \text{ when } n_j \text{ is even} . \end{aligned} \quad (4)$$

Like the mean, the median assumes that all observed results (including outlier values) are symmetrically distributed about the μ of the majority population. The median is much less sensitive than the mean to the presence of a modest number of outliers. For truly normal data without outliers, the median is about 20 % less efficient than the mean.

3.1.3 *Shorth*. The shorth (“shortest half”) is the value that is in the center of the most compact half of the reported results. It is the univariate analogue of the multivariate minimum volume ellipsoid estimator and has been used in at least one non-chemical KC.^{17,18} The value of the shorth is generally similar to the median, except that it does not require that outliers are symmetrically distributed about the μ of the majority population and it is considerably less efficient than the median when outliers are either not present or are symmetrically distributed. Although it has no closed form, the shorth is readily calculated from rank-ordered data

$$\text{Shorth}_j = \frac{x_{k,j} + x_{k+m,j}}{2} \quad (5)$$

where m is $n_j/2$ when n_j is even and $\text{int}(n_j/2)+1$ when n_j is odd and k is the index in the range 1 to n_j-m that minimizes $x_{k+m,j} - x_{k,j}$. For some data, two or more values for k may yield the same minimum difference. In such cases, a “grand shorth” can be defined as the average of the individual estimates.

3.1.4 *A15*. The A15 is a location metric that achieves robustness by truncating “outlier” values to be no more than a set distance from the central location. It assumes that outliers are symmetrically distributed about the μ of the majority population. It has essentially the same efficiency as the mean in the absence of outliers. The A15 does not have a closed form but is readily evaluated via iteration using the rule

$$\text{A15}_j = \sum_i^{n_j} \frac{z_{ij}}{n_j} \quad (6)$$

$$z_{ij} = x_{ij} \text{ if } |x_{ij} - \text{A15}_j| < c * \text{MADE}_j \text{ else } z_{ij} = x_{ij} + \text{sign}(x_{ij} - \text{A15}_j) * c * \text{MADE}_j$$

where MADE_j is a robust estimate of the standard deviation of the majority population calculated from the median absolute deviation of the x_{ij} from the median and c is an empirical constant (generally set to 1.5) that defines where “outlierness” begins in terms of number of MADE_j . Freeware spreadsheet software is available for the A15 estimate.¹⁹

3.1.5 *H15*. The H15 (also known as Huber’s Estimate 2) is very similar to the A15 with the exception that the standard deviation of the majority population is iteratively estimated simultaneously with the H15. The H15 thus makes more complete use of the data at the expense of somewhat greater computational complexity. The H15 has been strongly recommended for use with interlaboratory results.^{20,21} The iterative solution for H15 (and its associated standard deviation) is based on

$$\text{H15}_j = \sum_i^{n_j} z_{ij} / n_j$$

$$z_{ij} = x_{ij} \text{ if } |x_{ij} - \text{H15}_j| < c * s_j \text{ else } z_{ij} = x_{ij} + \text{sign}(x_{ij} - \text{H15}_j) * c * s_j \quad (7)$$

$$s_j = \beta \sqrt{\sum_i^{n_j} \frac{(z_{ij} - \text{H15}_j)^2}{n_j - 1}}$$

where β is a function of c that adjusts for the distortion introduced by the data truncation. Freeware spreadsheet software is available for the H15 estimate.

3.1.6 $L1\frac{1}{2}$. The $L1\frac{1}{2}$ is a member of the class of “least power” location metrics recently proposed for use with KC and other interlaboratory data.²² Given that the mean is the $L2$ metric and (when n_j is odd) the median is the $L1$ metric, $L1\frac{1}{2}$ has properties roughly “mid way” between the mean and the median. While not having a closed form, $L1\frac{1}{2}$ is readily solved by direct minimization

$$L_{1\frac{1}{2}j} = y \text{ where } \left(\sum_i^{n_j} |x_{ij} - y|^{\frac{3}{2}} \right)^{\frac{2}{3}} \text{ has minimum value .} \quad (8)$$

3.2 Uncertainty Weighting

The most common use of reported MU is in a weighting function that gives each of the x_{ij} more or less influence in a particular calculation based at least partly on the $U_{68}(x_{ij})$. Many different weighting functions have been proposed for use with interlaboratory studies. The following two variants have been used in the analysis of CCQM KCs.

3.2.1 *WtU*. The most commonly used weighted location metric is the simple “inverse-square” weighted mean. It is easily evaluated in closed form

$$\text{WtU}_j = \frac{\sum_i^{n_j} (w_{ij} x_{ij})}{\sum_i^{n_j} w_{ij}}; \quad w_{ij} = \frac{1}{U_{68}^2(x_{ij})} . \quad (9)$$

As a given $U_{68}(x_{ij})$ becomes relatively small, regardless of the x_{ij} value, the x_{ij} becomes relatively influential; as the $U_{68}(x_{ij})$ becomes relatively large, the x_{ij} becomes relatively inconsequential. The robustness of the WtU thus depends upon there being a strong positive association between “outlier” x_{ij} and relatively large $U_{68}(x_{ij})$. When all of the $U_{68}(x_{ij})$ are equal, the WtU reduces to the simple mean and will have the same efficiency as the mean. When the $U_{68}(x_{ij})$ are not equal, the efficiency of the WtU is related in a complex fashion to the veracity of the $U_{68}(x_{ij})$.

3.2.2 *WtMP*. Mandel-Paule weighting is an early version of what is now a large group of weighted location metrics that enforce a limit to the influence of unrealistically small $U_{68}(x_{ij})$.²³ It has recently been shown to be an approximate maximum likelihood location estimate.²⁴ The WtMP does not have a closed form but is solved numerically by iteration

$$\text{WtMP}_j = \frac{\sum_i^{n_j} (w_{ij} x_{ij})}{\sum_i^{n_j} w_{ij}} \quad (10)$$

$$w_{ij} = \frac{1}{U_{68}^2(x_{ij}) + s_j} \quad \text{where } \sum_i^{n_j} \frac{(x_{ij} - \text{WtMP}_j)^2}{U_{68}^2(x_{ij}) + s_j} = n_j - 1$$

where s_j can be interpreted as an estimate of the standard deviation of the combined data. The WtMP is about equally influenced by all x_{ij} with $U_{68}(x_{ij})$ smaller than s_j while being relatively little influenced by x_{ij} with $U_{68}(x_{ij})$ much bigger than s_j . Like the WtU, WtMP reduces to the mean when all of the $U_{68}(x_{ij})$ are equal and both the efficiency and robustness are related to the veracity of the $U_{68}(x_{ij})$ when the $U_{68}(x_{ij})$ are not equal.

3.3 Kernel Density Bootstrap Resampling

Bootstrap resampling has become a widely used method for estimating the value and variability of summary estimates when the observed data are not well approximated as belonging to one of the well-studied PDFs.²⁵ Bootstrap estimates are derived by analyzing a large number of “pseudo” datasets that are constructed by randomly and repeatedly drawing values from the observed data. What constitutes a “large number” is situation-specific, but is seldom less than a few hundred and may run into the millions. A “total median” location metric based upon the bootstrap concept has been proposed for use with KC data.

A modified bootstrap method that uses MU has been used with multi-source chemical data; rather than resampling just from the x_{ij} , the pseudo-data are obtained by random sampling from the $N(x_{ij}, U_{68}(x_{ij}))$ kernel densities.²⁶ That is, each resampling is accomplished by generating a random but characteristic value from a randomly chosen kernel. This kernel-density bootstrap reduces to the traditional bootstrap when the $U_{68}(x_{ij})$ values are asserted to be zero.

Since pseudo-data sets are randomly generated, bootstrap calculations are never “exact”. Two successive bootstrap estimates of the same value for the same data are likely to differ slightly, with the extent of agreement expected to improve as the number of pseudo-data sets evaluated is increased. There are also many different ways to summarize bootstrap results. While any or all of the traditional location metrics could be evaluated using the kernel-density bootstrap, only two such metrics are examined here.

3.3.1 BSmean. The BSmean is defined, for this study, as the mean of bootstrap means for pseudo-data sets of size n_j

$$\text{BSmean}_j = \frac{\sum_1^{NBS} \left(\sum_{i=1}^{n_j} R(x_{ij}, U_{68}(x_{ij})) / n_j \right)}{NBS} \quad (11)$$

where $R(x_{ij}, U_{68}(x_{ij}))$ is a random number drawn from a normal distribution having mean x_{ij} and standard deviation $U_{68}(x_{ij})$ and NBS is the number of pseudo-data sets generated. For this report, $NBS = 10,000$.

3.3.2 BSmedian. BSmedian is defined, for this study, as the median of bootstrap medians for pseudo-data sets of size n_j

$$\text{BSmedian}_j = \text{Median}_1^{NBS} \left(\text{Median}_{i=1}^{n_j} R(x_{ij}, U_{68}(x_{ij})) \right). \quad (12)$$

The same 10000 pseudo-data sets generated for the BSmean were used to estimate BSmedian.

3.4 Mixture Model Probability Density Functions

The MM-PDF of a combined set of results represents “all” of the information provided by the reported expected values and their associated MUs. Visualizing the MM-PDF at the margin of an otherwise routine dot-and-bar graph helps to establish the shape and dispersion of the combined results (Figure 2). MM-PDFs have been proposed as a way to empirically establish 95 % confidence intervals on traditional location estimates for KC results. Several different MM-PDF metrics have been proposed for use with CCQM KC

data. Unlike the kernel-density bootstrap estimates, the proposed MM-PDF metrics are in principle “exact” for a given set of combined results.

3.4.1 MMmode. The MMmode is a direct analogue of the mode, the most common value in a set of discrete values. It is the location at which the MM-PDF has greatest density; i.e., where the largest value is:

$$\text{MMmode}_j = y \text{ where } \sum_{i=1}^{n_j} \varphi(y, x_{ij}, U_{68}(x_{ij})) \text{ has maximum value .} \quad (13)$$

In practice, the MMmode is interpolated from the MM-PDF as evaluated on a reasonably dense, equally-spaced grid.

3.4.2 MMmedian. The MMmedian is a direct analogue of the median. It is the location which divides the MM-PDF into two sections of equal area.

$$\text{MMmedian}_j = Y \text{ where } \frac{\sum_{i=1}^{n_j} \int_{-\infty}^Y \varphi(y, x_{ij}, U_{68}(x_{ij})) dy}{n_j} = 0.5 \quad (14)$$

where $\varphi(y, x_{ij}, U_{68}(x_{ij}))$ is the probability density at y for a $N(x_{ij}, U_{68}(x_{ij}))$ kernel of unit area. In practice, the MMmedian is interpolated from the cumulative area of the MM-PDF as evaluated on a reasonably dense, equally-spaced grid.

3.4.3 MMsh/mid. The MMsh/mid is a direct analogue of the shorth (described above). It is the half-range location between the endpoints, YL and YH , of the shortest interval that contains one-half of the MM-PDF area

$$\text{MMsh/mid}_j = \frac{YL + YH}{2} \text{ for } \min(YH - YL) \text{ where } \frac{\sum_{i=1}^{n_j} \int_{YL}^{YH} \varphi(y, x_{ij}, U_{68}(x_{ij})) dy}{n_j} = 0.5 . \quad (15)$$

Like the MMmedian, these endpoints are in practice interpolated from the cumulative area of the MM-PDF as evaluated on a reasonably dense, equally-spaced grid. As with the shorth, it is possible that two or more intervals may be equally compact and may, in principle, be similarly made unique by averaging the multiple solutions.

3.4.4 MMsh/med. The MMsh/med is very similar to the MMsh/mid but makes more complete use of the MM-PDF by finding the location of the half-area between the YL and YH endpoints

$$\text{MMsh/med}_j = Y \text{ where } \frac{\sum_{i=1}^{n_j} \int_{YL}^Y \varphi(y, x_{ij}, U_{68}(x_{ij})) dy}{n_j} = 0.25 . \quad (16)$$

5 RESULTS AND DISCUSSION

5.1 Uncertainty Versus Bias

Weighting values (assigning more or less influence) on the basis of the magnitude of the associated MU implicitly asserts that the smaller the MU, the better the data; that is, that MU and measurement absolute-bias are positively correlated. Assuming that (1) the RVs stated in the Final Reports of the publicly available CCQM data sets are reasonable approximations to a “true” value for the measurands and (2) that all of the reported measurements can be validly interpreted as $N(x_{ij}, U_{68}(x_{ij}))$ kernels, it is possible to test this assertion.

The magnitude of measurement bias of a particular result is estimated as the absolute value of the difference between the expected result and the RV

$$\text{Bias}_{ij} = |x_{ij} - \text{RV}_j| \quad (17)$$

where RV_j is the assigned RV for the j^{th} measurand. While these bias estimates can be directly compared to the $U_{68}(x_{ij})$ within each data set, none of the CCQM data sets are sufficiently large to make such comparisons very informative. However, it is possible to compare the bias and MU estimates for all data sets at the same time by normalizing the bias and MU estimates to have a common scale

$$\text{NormalizedBias}_{ij} = \frac{|x_{ij} - \text{RV}_j|}{s_j}; \quad \text{NormalizedMU}_{ij} = \frac{U_{68}(x_{ij})}{s_j} \quad (18)$$

where s_j is a measure of the dispersion of the dataset. While there are many possible ways to estimate s_j , for this study it is defined from the central 50 % of the area of each PDF_j . This and related PDF-based dispersion metrics estimate the total variance of the combined data.

Figure 3 displays all 1294 of the normalized {Bias, MU} pairs for the currently available CCQM KC and pilot studies. The four symbols code the measurand type: ● denotes pH, □ gases, + organic, and × inorganic. All values larger than $4.0 s_j$ are plotted at the right or top margins. There is little if any evidence for any strong relationship between bias and MU in these data.

The data do, however, speak to the veracity of the CCQM MU estimates. The vertical line at normalized bias of $2.0 s_j$ is a routine threshold for dividing values that are apparently unbiased (to the left of the line) from those that may be (to the right): by this metric, over 90 % of the CCQM measurements agree well with the assigned RV. The thick line of slope 0.5 is the threshold separating measurements that include the RV within their 95 % confidence (above the line) from those that do not (below): over 72 % of the CCQM measurements reported MU at least as large as appropriate. The thin line of slope 0.5 is an approximate division between apparently complete (below) and potentially over-estimated (above) MU for the relatively unbiased measurements: by this metric, MU is over-estimated for only about 7 % of the measurements. In contrast, there is evidence for significant but unrecognized bias in about 8 % of all measurements.

5.2 Reference Values Vs Location Metrics

The majority of the CCQM measurement values agree well with the assigned RVs and most MU estimates appear realistic. However, given the number and diversity of the

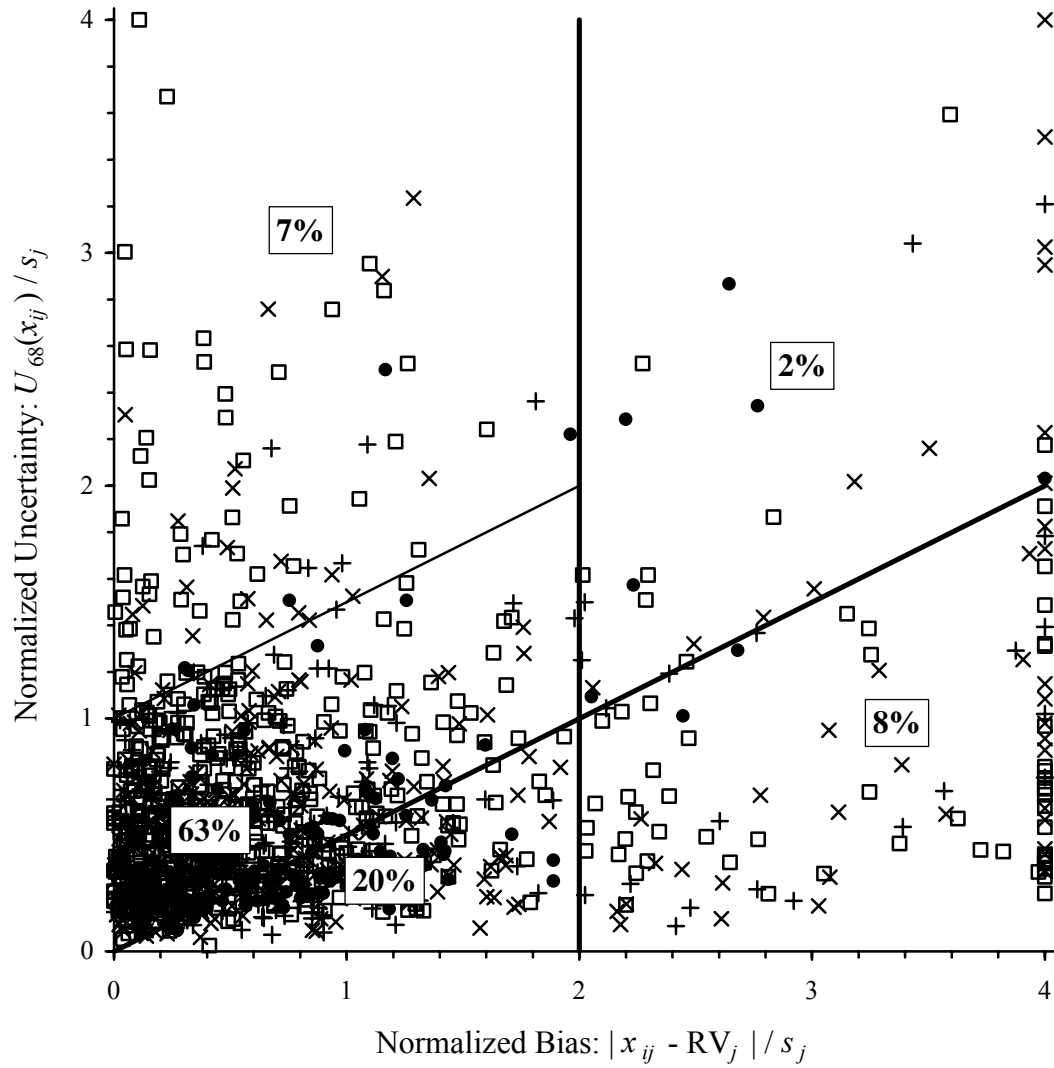


Figure 3 Measurement Uncertainty as a Function of Measurement Bias

measurands, the modest number of “outlier” values (the 10 % relatively biased), and both the apparently under- (28 %) and over-estimated (7 %) “outlier” MUs, these data should challenge any metric used to summarize the location of each set of combined results. Comparison of the RVs assigned by expert evaluation to the estimates produced by unguided application of the various metrics thus should help characterize some of the properties of the metrics.

5.2.1 Evaluation. The expected location of each of the 150 publicly available CCQM data sets was estimated with all 14 location metrics described above using spreadsheet software. The redundancy in the resulting 15 row (14 location estimates plus the assigned RV) by 150 column (measurands) data matrix was removed using principal components analysis. The resulting 15 estimate by 14 abstract-factor matrix captures 100 % of the covariance information contained in all 2250 location estimates.²⁷

Table 3 lists the scores for the first 10 factors. The variance explained by each n^{th} abstract factor is indicated by its eigenvalue, λ_n ; the proportion of the total variance, $\%Var_n$, explained by the factor is here $100 \cdot \lambda / 2250$; and cumulative proportion of the variance explained by the n largest factors is given by $\%CumVar_n$.

Table 3 Scores (c_{mn}) for the 10 Most Significant Factors of the Location Estimates

Metric	Abstract Factors									
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
RV	-3.9	-6.9	14.0	-0.7	1.8	0.7	-1.0	0.1	-0.1	0.1
Mean	12.0	0.1	-1.1	-1.5	0.6	-2.1	-4.7	-0.9	-2.0	-1.7
Median	-1.9	2.8	0.2	-2.5	-2.5	-1.5	2.3	4.2	-2.1	1.2
Shorth	-11.0	7.1	-1.3	-8.0	7.6	-1.1	0.4	-0.8	0.7	-0.2
A15	3.0	1.9	0.7	-2.9	-4.5	-0.7	1.8	-0.8	0.5	0.8
H15	5.0	1.2	0.8	-2.9	-5.0	-0.5	0.7	-3.1	1.7	0.9
L1½	4.0	1.6	0.0	-2.2	-2.4	-1.4	-0.1	0.2	-1.1	-0.3
WtU	0.7	-12.8	-5.7	-1.2	2.9	1.8	4.2	-1.3	-1.4	0.2
WtMP	6.7	-5.8	-3.8	-1.9	1.3	0.8	-3.6	3.1	3.1	0.4
BSmean	14.9	5.9	1.3	6.7	6.1	0.5	1.1	-0.4	-0.2	1.4
BSmedian	0.3	3.5	1.1	4.3	-0.7	2.4	4.5	1.0	1.5	-2.2
MMmode	-11.1	-2.7	-1.9	7.3	-0.2	-7.4	-0.6	-0.5	0.7	0.2
MMmedian	-0.6	0.1	-0.5	-0.4	-2.4	0.4	0.3	0.3	-0.1	-1.5
MMsh/mid	-8.6	2.6	-1.9	2.6	-1.8	5.6	-3.3	-0.7	-0.7	0.9
MMsh/med	-9.4	1.3	-1.9	3.2	-0.8	2.6	-2.1	-0.3	-0.6	0.0
λ	884.8	372.5	262.5	233.5	176.4	113.7	100.7	42.7	28.4	16.2
%Var	39.3	16.6	11.7	10.4	7.8	5.1	4.5	1.9	1.3	0.7
%CumVar	39.3	55.9	67.5	77.9	85.8	90.8	95.3	97.2	98.5	99.2

5.2.2 *Comparison of Metrics to Reference Values.* Location metrics that respond similarly to the various challenges presented in the evaluation of the 150 measurands will have similar score coefficients, c_{mn} , where m is the metric (row) index and n is the factor (column) index. Close examination of Table 3 indicates that the pattern of c_{mn} for MMsh/mid and MMsh/med are indeed quite similar.

It is, however, much more convenient to quantitatively summarize the similarities of the different metrics with the Euclidean score distance, $D_{p,q}$:

$$D_{p,q} = \sqrt{\sum_{n=1}^{14} (c_{pn} - c_{qn})^2} . \quad (19)$$

where p and q are the row designations for two metrics. Should the two metrics have exactly the same properties, the value of $D_{p,q}$ will be 0.0; the value of $D_{p,q}$ will increase as the similarity between the metrics' behavior decreases. For reference, the score distance between MMsh/mid and MMsh/med, $D_{14,15}$, is 4.7.

Figure 3 displays the $D_{p,q}$ for the 14 location metrics examined (rows 2 to 15 of Table 3) relative to the RV (row 1). The metrics are sorted in order of decreasing similarity to the RV. These distances suggest two broad conclusions. First, none of the 14 metrics gives location estimates that are closely related to the RV: unguided statistics cannot replace expert evaluation. Second, of the metrics evaluated, the MMmedian is distinctly the most similar to the RV: chemical MU estimates can provide quantitatively useful information.

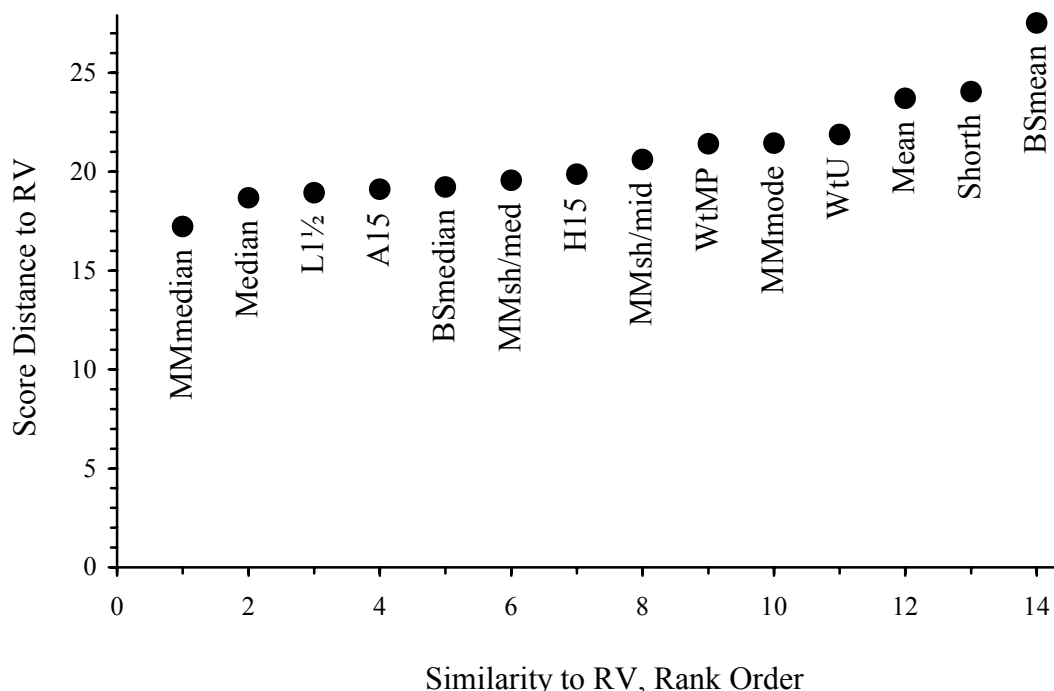


Figure 3 *Factor-Score Distance of Unguided Location Metrics to Reference Values*

5.2.2 *Comparison of Metrics to the MMmedian.* Figure 4 displays the $D_{p,q}$ of the RV and the various location metrics to the MMmedian. These values confirm that while the MMmedian estimates may be the most similar of the metrics evaluated to the RVs, they are much more closely related to those from most of the other metrics than to the RVs. The MMmedian appears to be most similar to the traditional metrics that are robust to symmetrically distributed “outlier” values (L1½, A15, Median, and H15) and least similar to the uncharacterized BSmean and the putatively robust but asymmetric Shorth. The behaviors of the other asymmetric metrics (MMsh/med, MMsh/mid, MMmode) are somewhat more similar than the Shorth to both the MMmedian and to the RVs, with the MMsh/med the most efficient of the asymmetric metrics for these basically symmetrically-distributed CCQM data.

The WtMP, WtU, and Mean metrics are about equally dissimilar to the robust estimates and to the RVs, confirming that measurement absolute-bias and MU are not sufficiently correlated in the CCQM data to identify relative bias from the relative MUs. The small difference between WtMP and WtU suggests that using MU to allocate influence is a basic fallacy rather than a deficiency in how such influence is allocated.

While the BSmedian estimates are roughly similar to those of the traditional robust metrics, they are not as close to the RV or the MMmedian as the median itself. In contrast, the BSmean estimates are the least similar to the MMmedian and the RVs of any metric. While chemical MU estimates provide information useful for combining results, as always the devil is in the details of how that information is used.

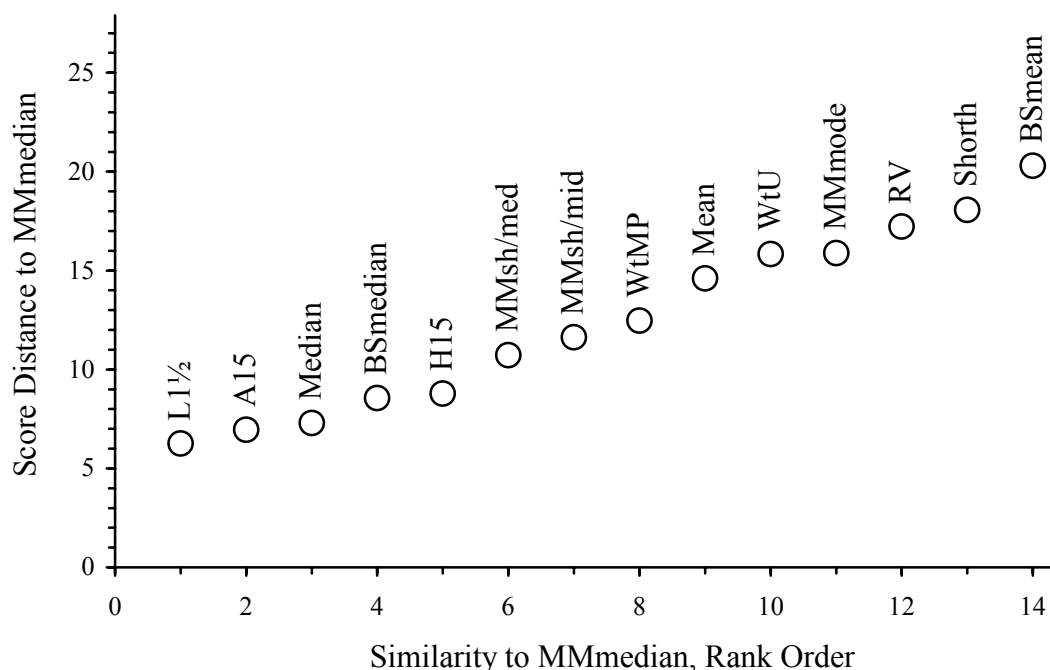


Figure 4 Factor-Score Distances to MMmedian

5 ACKNOWLEDGEMENT

I thank Reenie M. Parris, Kenneth W. Pratt, Michele M. Schantz, and Katherine E. Sharpless of NIST for their questions and gentle corrections; Aristides T. Hatjimihail of the Hellenic Complex Systems Laboratory for his encouragement; Richard G. Brereton of the University of Bristol for his Multivariate Analysis spreadsheet freeware, and Steven L.R. Ellison of LGC for his RobStat spreadsheet freeware and for his deep insights into the use and abuse of chemical MU.

6 REFERENCES

- 1 Eurachem. The fitness for purpose of analytical methods. A laboratory guide to method validation and related topics. 1998, www.eurachem.ul.pt/guides/valid.pdf
- 2 Eurachem. Quantifying uncertainty in analytical measurement, 2nd Ed. 2000, www.eurachem.ul.pt/guides/QUAM2000-1.pdf
- 3 G.E. O'Donnell, D.B. Hibbert. *Analyst*, 2005, **139**, 721.
- 4 CLSI. Expression of Uncertainty of Measurement in Clinical Laboratory Medicine (C51), *product in development*. www.clsi.org
- 5 M.N. Leger, L. Vega-Montoto, P.D. Wentzell. *Chemometrics Intell. Lab. Syst.*, 2005, **77**, 181.
- 6 M.S. Reis, P.M. Saraiva. *AIChE J.*, 2005, **51**, 3007.
- 7 P.J. Lowthian, M. Thompson. *Analyst*, 2002, **127**, 1359.
- 8 M.G. Cox. NPL Report CISE 42/99. 1999. www.npl.co.uk/ssfm/download/nplreports.html
- 9 P. Ciarlina, M.G. Cox, F. Pavese. G. Regoliosi. *Metrologia*, 2004, **41**, 116.

- 10 D.L. Duewer, J. Brown Thomas, M.C. Kline, W.A. MacCrehan, R. Schaffer, K.E. Sharpless, W.E. May, J.A. Crowell. *Anal. Chem.* 1997, **69**, 1406.
- 11 BIPM. kcdb.bipm.org/AppendixD/default.asp.
- 12 BIPM. www.bipm.fr/en/committees/cc/ccqm/pilot_cc.html
- 13 D.L. Duewer. Working document CCQM/04-15, BIPM, 2004. www.bipm.info/cc/CCQM/Allowed/10/CCQM04-15.pdf
- 14 C. Croux, G. Haesbroeck. *J. Nonparametr. Stat.*, 2002, **14**, 295.
- 15 J.W. Müller. *J. Res. Nat. Inst. Stds. Technol.* 2000, **105**, 551.
- 16 M.G. Cox. *Metrologia* 2002, **39**, 589.
- 17 P.J. Rousseeuw. In: W. Grossman, G. Pflug, I. Nincze, W. Wetz (eds.), *Mathematical Statistics and Applications*, 1985, 283-297. Reidel, Dordrecht, The Netherlands.
- 18 A.H. Rose, C.-M. Wang, S.D. Byer, *J. Res. Nat. Inst. Stds. Technol.* 2000, **105**, 839.
- 19 Analytical Methods Committee. RobStat.xla. 2002. www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/RobustStatistics.asp
- 20 Analytical Methods Committee. AMC Technical Brief 6. 2002. www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp
- 21 Analytical Methods Committee. *Analyst* 1989, **114**, 1693.
- 22 F. Pennechi, L. Callegaro. *Metrologia*, 2006, **43** 213.
- 23 R.C. Paule, J.Mandel. *J. Res. Nat. Bur. Stds.* 1982, **87**, 377.
- 24 A.L. Rukhin, B.J. Biggerstaff, M.G. Vangel. *J. Stat. Plan. Infer.* 2000, **83**, 319.
- 25 P. Diaconis, B. Efron. *Sci. Am.* 1983, **248**, 116.
- 26 D.L. Duewer, B.R. Kowalski, J.L. Fasching. *Anal. Chem.* 1976, **48**, 2002.
- 27 R.G. Brereton. *Chemometrics : Data Analysis for the Laboratory and Chemical Plant*, Wiley, Chichester, 2003.