Accred Qual Assur (2007) 12:175–187
DOI 10.1007/s00769-006-0245-2

PRACTITIONER'S REPORT

Description and results of the NIST/NOAA 2005 Interlaboratory Comparison Exercise for Trace Elements in Marine Mammals

Steven J. Christopher · Rebecca S. Pugh · Michael B. Ellisor · Elizabeth A. Mackey · Rabia O. Spatz · Barbara J. Porter · Kathie J. Bealer · John R. Kucklick · Teri K. Rowles · Paul R. Becker

Received: 22 November 2006 / Accepted: 16 December 2006 / Published online: 24 January 2007

Abstract The National Institute of Standards and Technology's (NIST) National Marine Analytical Quality Assurance Program (NMAQAP) is dedicated to improving the quality of analytical measurements of trace elements, organic contaminants and emerging compounds of concern in marine and environmental systems, through various quality assurance mechanisms, including analytical method development and value assignment, quality assurance materials production, cryogenic marine specimen archival and the coordination of interlaboratory comparison exercises. This report focusses on the description and results of the 2005 Interlaboratory Comparison Exercise for Trace Elements in Marine Mammals. This program is co-sponsored by the National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Office of Protected Resources, specifically, the Marine Mammal Health and Stranding Response Program. Two quality control materials derived from fresh-frozen marine mammal livers were produced and characterised at the NIST and were then distributed to over 30 laboratories. A maximum likelihood solution model was used to assign consensus data that served as a benchmark for comparison, and a series of group metrics were generated to assist the laboratories with the interpretation of performance and analytical assessment.

Keywords Consensus mean · Interlaboratory comparison exercise · Maximum likelihood · Marine mammals · Trace elements

Introduction

It is important to underpin the measurement accuracy of results from laboratories that perform marine environmental analyses. The ability to accurately determine trace analytes in a wide range of marine sample types is required to assess their impact on human and animal health and nutrition, provide temporal "snapshots" of marine environmental quality and to identify global, regional and point sources that release contaminants into the atmosphere and coastal ecosystems. Critical reference standards are often not available for this niche analytical community, especially reference materials derived from protected marine species. This limitation may lead to species management decisions that are based on ambiguous analytical results, which can have significant environmental, economic and health consequences.

The National Institute of Standards and Technology (NIST) helps benchmark and improve the quality of

Presented at "BERM-10", April 2006, Charleston, SC, USA.

S. J. Christopher (✉) · R. S. Pugh · M. B. Ellisor ·
K. J. Bealer · J. R. Kucklick · P. R. Becker
Analytical Chemistry Division,
Hollings Marine Laboratory,
National Institute of Standards and Technology,
Charleston, SC 29412, USA
e-mail: steven.christopher@nist.gov

E. A. Mackey · R. O. Spatz · B. J. Porter
Analytical Chemistry Division,
National Institute of Standards and Technology,
Gaithersburg, MD 20899, USA

T. K. Rowles
National Oceanic and Atmospheric Administration,
National Marine Fisheries Service,
Office of Protected Resources, Silver Spring,
MD 20910, USA

analytical data gathered on the marine environment by administering annual interlaboratory comparison exercises through several programs, including the National Oceanic and Atmospheric Administration's (NOAA) National Status and Trends Program and the National Marine Analytical Quality Assurance Program (NMAQAP), which is supported by the NOAA National Marine Fisheries Service, Office of Protected Resources (NOAA/NMFS), specifically, the Marine Mammal Health and Stranding Response Program. The NIST activities that are focussed on marine specimen banking, quality assurance and interlaboratory comparison exercises for these programs have been summarised by Becker et al. in 1997 [1]. The NIST produces quality control and reference materials that are distributed in annual interlaboratory comparison exercises, organises and coordinates the exercises, and performs baseline analytical measurements on marine samples collected and stored in the NIST National Biomonitoring Specimen Bank (NBSB) in support of this program. Participation in the interlaboratory comparison exercise program is currently offered at no cost to interested participants. Operating these programs in concert with the NIST Chemical Science and Technology Laboratory's (CSTL) Standard Reference Material (SRM) value assignment and analytical method development activities has helped the NIST to establish a comprehensive chemical measurement and quality assurance infrastructure to address marine-related chemical measurement issues.

The NMAQAP includes both an organic constituent and a trace element component. The diversity of the 33 participating institutions represented in the trace element component testifies that this interlaboratory comparison exercise extends beyond the scope of the NMAQAP to the trace element analytical community as a whole, including domestic and international health, environmental and diagnostic laboratories, academic institutions, contract and industrial laboratories, and government agencies. The intent of this paper is to describe the design of the exercise and the analysis of the trace element results, and to discuss the relevant outputs that ultimately allow the participants to assess their performance relative to their peers and those laboratories operating in the field of marine environmental research that encompasses measurements of trace elements.

This year (2005) marks the fourth iteration of the interlaboratory comparison exercise. Participants were asked to perform measurements for a suite of 15 analytes (Ag, As, Cd, Co, Cs, Cu, Fe, Hg, Mn, Mo, Rb, Se, Sn, V and Zn) in two NIST quality control materials: a pygmy sperm whale liver homogenate, QC03LH3, and a white-sided dolphin liver homogenate, QC04LH4. These samples are fresh-frozen quality control materials that were cryogenically pulverised, homogenised and bottled using established techniques [2]. Herein, the key results of the exercise and the statistical tools used for the data evaluation are presented. Consensus data were generated using the Rukhin–Vangel maximum likelihood (ML) estimation model [3], which uses weighted means statistics and considers both within- and between-laboratory variances. This data is compared to the data generated using robust statistics to assess the efficacy of the exercise design and consensus mean estimator model as applied to trace element data. The International Union of Pure and Applied Chemistry (IUPAC) guidelines were implemented to evaluate laboratory performance through the use of z- and p-scores [4], which provide a mechanism to assess the comparability of data produced by the participating laboratories. Group metrics of performance are presented and, finally, laboratory biasses are also evaluated graphically through the use of Youden diagrams [5].

Exercise details

Description of test materials

Pygmy sperm whale (*Kogia breviceps*) liver homogenate (QC03LH3) served as the control standard for the interlaboratory comparison exercise, while white-sided dolphin (*Lagenorhynchus acutus*) liver homogenate (QC04LH4) served as the unknown. QC03LH3 was prepared from the liver of a single live-stranded animal found at Sullivan's Island, Charleston County, SC, USA, on 10th August 1994. The collection effort was spearheaded by personnel at the NOAA National Ocean Service's Center for Coastal Environmental Health and Biomolecular Research, NOAA Fisheries and the SC Department of Natural Resources in Charleston, SC, USA. The material was donated to the NIST through the vehicle of the National Marine Mammal Tissue Bank, a component of the NMAQAP. The white-sided dolphin liver that was used to prepare QC04LH4 was donated by personnel at the New England Aquarium. All tissues were cryogenically pulverised, homogenised and bottled under ISO class 7 and class 5 clean room conditions to provide fresh-frozen, powder-like materials.

Exercise participation requirements and target analytes

The list of participating institutions is presented in Table 1. These laboratories include domestic and

Table 1 List of participating institutions

Participating institution	Country
Applied Speciation and Consulting LLC	USA
Australian Nuclear Science and Technology Organization	Australia
Brooks Rand LLC	USA
Cantest Limited	Canada
Centre for Environment, Fisheries and Aquaculture Science Burnham Laboratory	United Kingdom
Centre For Public Health Sciences, Queensland Health Scientific Services	Australia
Chungnam University Department of Chemistry	S. Korea
University of Massachusetts Department of Chemistry	USA
University of Canberra Ecochemistry Laboratory	Australia
University of Connecticut Environmental Research Institute	USA
U.S. Department of Agriculture Food Composition Laboratory Beltsville Human Nutrition Research Center	USA
Frontier Geosciences Incorporated	USA
Galab Laboratories	Germany
GBC Scientific Equipment	Australia
GKSS Research Center Institute for Coastal Research Department for Marine Bioanalytical Chemistry	Germany
Health Canada—Radiation Protection Bureau	Canada
Hercules Incorporated	USA
Hewlett Packard Company	USA
Institute of Chemistry—Analytical Chemistry Karl-Franzens University Graz	Austria
Izmir Yuksek Teknoloji Enstitusu	Turkey
Kinectrics Incorporated	Canada
Midwest Research Institute Florida Division	USA
National Measurement Institute, Pymble	Australia
National Measurement Institute, South Melbourne	Australia
Ontario Ministry of Environment Laboratory Services Branch	Canada
Politechnika Poznanska Department of Analytical Chemistry	Poland
Sawyer Environmental Research Center University of Maine	USA
Spectrometry Application Laboratory	Italy
Trace Element Research Laboratory Texas A&M University	USA
University of California Los Angeles Inductively Coupled Plasma Facility, Department of Chemistry and Biochemistry	USA
Ultra-Trace Analyses Aquitaine (UT2A)	France
Universitaet Hohenheim Landesanstalt Fuer Landwirtschaftliche Chemie	Germany
Université De La Rochelle Centre Commun D’analyses	France
University of California, Davis Plant Science Department	USA
University of Maryland Eastern Shore George Washington Carver Science Building	USA
University of Pennsylvania School of Veterinary Medicine New Bolton Center	USA

international public, private and academic institutions. The participating institutions were each sent glass jars containing approximately 8–10 g of each of the above frozen materials using liquid nitrogen (LN_2) vapour or dry ice shippers. Typically, the LN_2 shippers were used for overseas shipments and the dry ice shippers were used for domestic shipments. Shipments of samples of the types used in the exercise are subject to permitting under the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) [6]. The NIST works with international laboratories on an as-needed basis to ensure that all of the appropriate documentation is in place to deliver the materials in an efficient manner that preserves sample integrity.

Several requirements were stipulated to the participants. They were asked to keep the samples frozen, preferably at –80°C, prior to analysis. The exercise directions required that participants: (1) analyse samples for elements (As, Cd, Cu, Fe, Hg, Mn, Mo, Rb, Se, Sn, V and Zn) using accepted in-house analytical procedures; (2) digest, process and analyse three aliquots of QC03LH3; and (3) digest, process and analyse five aliquots of QC04LH4. The submission of wet mass fraction data occurred by electronic mail and only the raw data submitted for the individual determinations were used by the NIST, which handled all of the statistical processing. Participants were not asked to submit expanded uncertainty data as defined by the International Organization for Standardization [7]. Thus, the uncertainties calculated from the raw datasets are based solely on laboratory repeatability measurements derived from the analysis of multiple aliquots of a processed sample, which comprises components of uncertainty, including method repeatability, instrumental measurement repeatability and sample heterogeneity.

Statistical methods

Outlier testing

The reported laboratory results for the control sample, QC03LH3, were used to determine potential gross outliers in the data. First, the measurement capability for each measurand was evaluated by comparing the results for the QC03LH3 control sample against established target values calculated using a composite of the NIST analytical techniques and the consensus data generated on this material during the 2003 interlaboratory comparison exercise when it was issued as an unknown. The NIST typically uses instrumental neutron activation analysis (INAA) [8, 9] and inductively coupled plasma mass or emission spectrometry as in-house techniques to evaluate our fresh-frozen QC materials. Established mass fraction values, expanded uncertainties and target reference ranges for outlier testing for elements in QC03LH3 are presented in Table 2. The reference data for QC03LH3 are derived from combining data from the aforementioned sources using the Type B on Bias (BOB) method [10], which produces an equally weighted mean from the independent group means and associated expanded uncertainty that includes components of within- and between-laboratory variances.

Table 2 Mass fraction values, expanded uncertainties ($U_{k=2}$) and target range outlier criteria for elements in the control sample, QC03LH3

Element	Mass fraction (mg/kg)	$U_{k=2}$ (mg/kg)	Target range ±20% (mg/kg)
Ag	0.088	0.007	0.070–0.106
As	0.398	0.019	0.318–0.478
Cd	5.94	0.21	4.75–7.13
Co	0.071	0.003	0.057–0.085
Cs	0.0079	0.0003	0.0063–0.0095
Cu	2.74	0.12	2.19–3.29
Fe	694	25	555–833
Hg	3.56	0.67	2.85–4.27
Mn	1.43	0.07	1.14–1.72
Mo	0.211	0.008	0.169–0.253
Rb	1.61	0.07	1.29-1.93
Se	7.87	0.88	6.30–9.44
Sn	0.094	0.019	0.075–0.113
V	0.0370	0.0168	0.0296–0.0444
Zn	21.15	0.97	16.92–25.38

The laboratories were asked to analyse three subsamples of QC03LH3. The data were defined as outliers for particular elements if the difference between the reported laboratory mean result for QC03LH3 and the mean of the QC03LH3 reference data differed by 20% or greater. Corresponding trace element mass fraction data for the unknown sample, QC04LH4, were considered as outliers, regardless of the degree of agreement between the reported result and the consensus mean value, if adequate performance on the control could not be demonstrated. Outlier data were not used in the determination of the consensus means for elements in the unknown sample. This gross outlier rejection protocol worked well to identify laboratory results that would distort the consensus mean of the unknown sample, QC04LH4, the metric used as a point of reference to assess each laboratory's performance. The data were also treated as outliers if the protocol was violated (a minority of instances); examples include not reporting control data or reporting only a single measurement for the unknown sample, which precluded the establishment of inverse variance laboratory weights—a constraint for the consensus mean processing algorithm applied.

Consensus mean calculations

There are many approaches used at the NIST to compute an estimate of a consensus mean and its associated uncertainty, based on using datasets from multiple laboratories and/or multiple analytical methods [11–14]. The consensus means determined in this exercise are based on the weighed mean of the individual laboratory means, and this weighted mean was calculated using an iterative ML solution model [3]. When choosing a model to estimate a consensus mean, several fundamental factors must be considered. For any given analyte, the number of individual measurements performed and reported may vary across the laboratories, as individual laboratories may follow their routine processes and protocols rather than explicit directions. Thus, a consensus mean estimator model should be able to handle unbalanced datasets. Moreover, the within-laboratory variances can differ across the laboratories—this could be a function of method or material. Finally, the number of laboratories will also influence the choice of method used to estimate the consensus mean. These factors should determine how to appropriately weight each laboratory or whether to treat all laboratories equally. The forthcoming discussion will help to illustrate these points.

Homoscedasticity plots (laboratory standard deviation versus reported laboratory mean concentration) were generated for each element in the unknown sample, QC04LH4. The plots are not included here, but the vertical scatter observed in the plots indicated that the variances across the laboratories were not equal; thus, the assumption of equal variances across the laboratories does not hold for the reported interlaboratory data. A consensus mean estimator model that is based on weighted means statistics may be more applicable than a simple ‘‘mean of means’’ model, where the estimate is an equally weighted mean that does not account for possible differences in within-laboratory variability.

Consensus data are often used to ‘‘grade’’ each participating laboratory based on the proximity of its data to the consensus value, for example, using z- and p-scores to measure congruence and relative laboratory repeatability, respectively, according to IUPAC guidelines [4], as performed in this exercise. Therefore, it is desirable to incorporate an outlier rejection scheme and to also provide a reasonable estimate of the confidence interval about the consensus mean that, if possible, incorporates both within- and between-laboratory variance. This allows each participating laboratory to consider the merit and quality of the consensus value estimate (often treated as the ‘‘true’’ value by the participants) as a point of reference. The distribution of the analyte data should always be considered as well, as most estimation models assume that the data will follow a normal distribution. Figure 1 gives example histograms and normal probability plots for the Se raw data submitted for QC04LH4. The histogram and normal probability plot in Fig. 1a indicate graphically that this particular dataset is non-normally distributed. Applying a Shapiro–Wilk test to the data corroborates the visual indications, i.e. p<0.0001 is lower than the 95% significance level for p (0.05), and non-normality can be assumed. The histogram and normal probability plot are regenerated in Fig. 1b after removing suspected outlier laboratories. Here, the results for the Shapiro–Wilk test yielded p=0.09, and normality can be assumed. The data shown here for Se are representative of the data for the other elements; thus, the assumption of normality is applicable to the data in this exercise with the caveat that outlier data (if left unaccounted for) could easily negate the ‘‘normality’’ of a dataset.

The Rukhin–Vangel ML model [3] used in this exercise addresses a number of items discussed above. The model chosen for computing the consensus mean estimates can handle unbalanced datasets and helps to de-emphasise laboratory means that possess large variances. The ML consensus mean algorithm is typically reserved for application to interlaboratory datasets larger than six. It is important again to make the distinction between the procedures used in this exercise and the more familiar ‘‘mean of means’’ procedure for calculating the consensus mean, where the latter approach necessarily weights each laboratory identically, regardless of its analytical repeatability. The assumptions with regards to weighting laboratory data for this exercise are that accuracy and precision are correlated, and the most precise data should, therefore, be weighted more heavily. These assumptions are historically based on interlaboratory comparison exercises that prescribe and hold the analytical methods and procedures constant, which is not the case for this exercise, but the intention here is to apply the model to the trace element data collected and compare results to data generated using robust statistics, rather than challenge the underpinning assumptions of the model. It is obvious that weighting laboratory data should only be considered if the distributed samples are homogeneous, so as not to confound laboratory repeatability. Although the model is simple, the ML equations are rather complicated in form and are not reproduced here. See Rukhin and Vangel [3] for a rigorous presentation of this model. A very short overview is included here to outline the procedures used to determine an ML estimate of the consensus mean. The ML solution used to estimate the consensus mean and its associated uncertainty is based on a one-way random effects analysis of variance (ANOVA) model that may be both unbalanced (i.e. the number of observations from each laboratory need not be equal) and heteroscedastic (i.e. the within-laboratory variances can be unequal):

$$x_{ij} = \mu + L_i + e_{ij}$$

where there are i=1,..., k laboratories and j=1,..., n_i observations for each laboratory. In this model, μ is the consensus mean, L_i is the lab effect and e_{ij} is the error term. The L_i are normally distributed as N(0, σ^2) and the e_{ij} are normally distributed as N(0, σ_i^2). Here, σ^2 and σ_i^2 represent the between-laboratory and within-laboratory variances, respectively. The Rukhin–Vangel paper outlines the likelihood function [3]. The ML estimate $\bar{x}$ of the consensus mean μ is solved using the following equation:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \gamma}{\sum_{i=1}^{k} \gamma} \tag{1}$$

where x_i is the reporting laboratory mean and the summation is from i=1 to k, where k is the number of

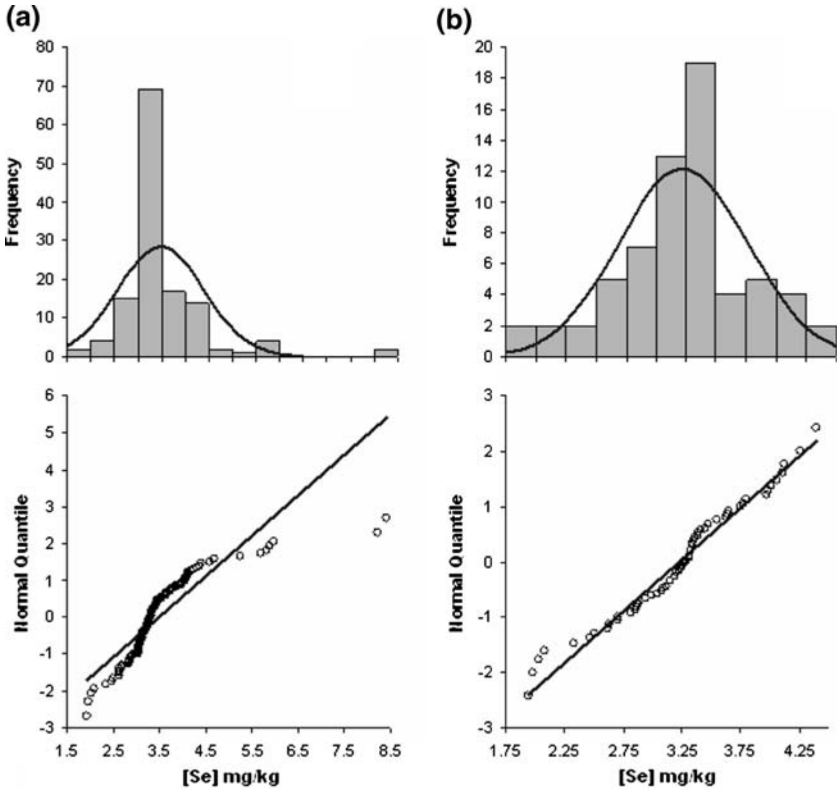

Fig. 1a, b Histogram and normal probability plots for selenium in QC04LH4 before (**a**) and after (**b**) the removal of outliers

laboratories. The quotients $\gamma = \frac{\sigma^2}{\left(\sigma^2 + \frac{\sigma_i^2}{n_i}\right)}$ are the weights to be assigned to each laboratory that contributes data to the consensus mean. These assigned weights are inversely proportional to the within-laboratory variances. The ML estimate $\hat{\sigma}^2$ of the between-laboratory variance, is solved using the following equation:

$$\hat{\sigma}^2 = \sum_{i=1}^{k} \frac{\gamma\left[(x_i - \bar{x})^2 + \frac{(n_i - 1)\frac{s_i^2}{n_i}}{1-\gamma}\right]}{N + k} \tag{2}$$

An ML estimate for each γ is determined numerically, through an iterative process [3], in order to solve the above equations. The standard uncertainty of the estimate of the consensus mean is then computed using the following formula:

$$\bar{x} \pm \frac{\sqrt{\sum_{i=1}^{k} \frac{(x_i - \bar{x})^2}{\left(\hat{\sigma}^2 + \frac{\sigma_i^2}{n_i}\right)^2}}}{\sum_{i=1}^{k} \frac{1}{\hat{\sigma}^2 + \frac{\sigma_i^2}{n_i}}} \tag{3}$$

where the summation is from i=1 to k and k is the number of laboratories. Finally, this standard uncertainty is multiplied by a coverage factor (k=2), and this expansion is expected to provide an approximate 95% level of confidence for all of the analytes evaluated.

Consensus data were also generated using robust statistics to effect a comparison and evaluate any potential advantages of using the somewhat complicated schemes for outlier rejection and weighting laboratory data to yield the consensus estimates. The reported laboratory data for all elements tested were subjected to a robust statistical procedure using median and median average deviation statistics. No prerequisites were placed on the data; the unknown (QC04LH4) data was used as received and no control material data or performance qualifiers were implemented to identify gross outliers or to help establish traceability. Instead, outlier testing for the unknown material utilised Hampel Scoring [15]. The Hampel Score is easily calculated by subtracting the median concentration value of the full dataset across the laboratories from the mean laboratory concentration, computing the absolute value of this quantity and dividing by the product 1.4826[median average deviation (MAD)] [15]. Data with HS>3 were flagged as outliers and a new median value was obtained (after the removal of outliers), which served as the consensus mean estimate. MAD and MADe, an estimate of the standard deviation based on MAD/0.674, was calculated using RobStat Software [16, 17], and a rough 95% confidence interval about the median consensus estimate was assigned by multiplying the MADe standard deviation estimate by a coverage factor of 1.96. The approach described necessarily assumes that the data are distributed normally after the removal of outliers.

Performance assessment tools

z- and *p*-scores

The *z*-score is a bias estimate calculated from the difference between the laboratory mean x_i and the consensus mean estimate divided by a target value (σ_{Target}) for the standard deviation:

$$z = \frac{x_i - \bar{x}}{\sigma_{\mathrm{Target}}} \tag{4}$$

The choice of σ_{Target} will be dependent on the data quality objectives of a particular quality assurance program. For this exercise, *z*-scores are calculated using a fixed fit for performance criterion $\sigma_{\mathrm{Target}}=\pm 10\%$ of the consensus mean. Using two examples, this performance criterion implies that, respectively for $z=\pm 1$ or $z=\pm 2$, the result is 10% or 20% higher (or lower) than the consensus mean. One should use *z*-scores to comment on congruence and not absolute concentration accuracy. *z*-scores have traditionally been divided into categories to assess the performance of each laboratory: $|z|\leq 2$ is satisfactory, $2\leq|z|\leq 3$ is questionable and $|z|\geq 3$ is unsatisfactory.

Using a ‘‘fixed’’ performance criterion offers a way for each laboratory to compare their performance on different samples and against other participating laboratories. It should be recognised that any particular laboratory might have a detection limit or analytical method deficiency for a particular analyte; thus, the acceptability of a particular laboratory's results should, in this exercise, be judged by the participants themselves in the context of the data quality needs and objectives of each particular program. The external repeatability of each laboratory for individual elements is assessed using a *p*-score (precision score), where laboratory repeatability (i.e. the coefficient of variation, CV) is normalised to an assigned target value for the coefficient of variation:

$$p = \frac{\mathrm{CV}_{\mathrm{Lab}}}{\mathrm{CV}_{\mathrm{Target}}} \tag{5}$$

The value for $\mathrm{CV}_{\mathrm{Target}}$ is fixed at 10% for this interlaboratory comparison exercise. Using an example, this value for $\mathrm{CV}_{\mathrm{Target}}$ implies that, for $p=0.5$, the laboratory repeatability is 5%. However, sample inhomogeneity is a limiting factor when evaluating intralaboratory or interlaboratory repeatability.

Youden diagrams

The Youden diagram [5] is a classic graphical tool used to evaluate laboratory bias when each laboratory has collected data on two similar materials. The Youden diagram is an effective means for comparing between- and within-laboratory variability and highlighting possible outliers. This graphical tool helps assess whether the laboratories in the study are behaving as a single population and can be used to provide information on the occurrence of indeterminate (random) and determinate (systematic) errors. A Youden diagram will exhibit a structureless ‘‘random shotgun pattern’’ about a point of reference [5] if all laboratories reside within a single population and indeterminate errors are dominant. Measurements appearing in the upper right and lower left quadrants of the diagram indicate, respectively, that a laboratory's measurements are consistently biassed high or low relative to measurements performed in other laboratories. Sources of such determinate errors include calibration errors, blank correction errors, analytical method errors such as analyte volatility (loss) and sample contamination, and matrix and spectral interferences. Measurements appearing in the upper left or lower right quadrants may indicate sample heterogeneity or sample-specific method bias.

The relative point of reference for the Youden diagrams used in this study was the intersection of the assigned reference value for the control material (QC03LH3) and the ML consensus mean value calculated for the unknown material (QC04LH4). An example Youden diagram for Cu is presented in Fig. 2. The normalised bias reference point (intersection at coordinates $x=1$ and $y=1$ in the example diagram) represents the best estimate of congruence for Cu tested in the two materials. A two-dimensional 95% confidence interval is cast about the point. Measurements from individual laboratories are normalised to the reference and consensus values described above so that they can all be compared against a common benchmark. In general, laboratories falling closer to the bias reference point in a Youden diagram demonstrate congruence.

Results and discussion

Survey of analytical methods

Table 3 displays the reported instrumental methods as a percentage of use for each element. The reported

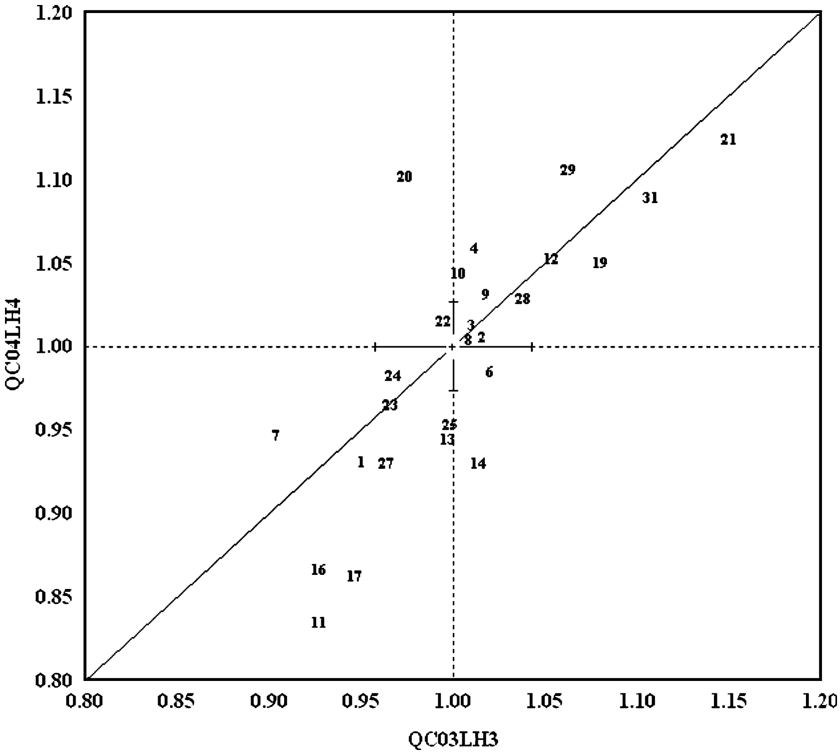

Fig. 2 Youden diagram for laboratory Cu measurements in the control (QC03LH3) and unknown (QC04LH4) samples

Table 3 Percentage use of reported instrumental methods as a function of element

Element	Unknown	Q-ICPMS	ICPMS	ICP-OES	INAA	TXRF	Other
Ag	25	46	17	8	4		
As	28	40	13	10	3	3	3
Cd	29	38	12	15	3		3
Co	27	42	15	12	4		
Cs	33	44	17	6			
Cu	29	34	11	14	3	6	3
Fe	25	31	13	25	3	3	
Hg	28	21	14	10	3	3	21
Mn	26	38	12	15	3	3	3
Mo	31	42	15	8	4		
Rb	27	41	14	9		9	
Se	24	36	13	10	3	7	7
Sn	22	46	18	9	5		
V	27	46	18	9			
Zn	29	32	12	18	3	6	

Unknown=unreported method; *Q-ICPMS*=quadrupole inductively coupled plasma mass spectrometry; *ICPMS*=sector-field and collision/reaction cell inductively coupled plasma mass spectrometry; *ICP-OES*=inductively coupled plasma optical emission spectrometry; *INAA*=instrumental neutron activation analysis; *TXRF*=total X-ray fluorescence; *Other*=cold vapour, hydride generation or electrothermal atomic absorption and fluorescence spectrometry

methods are heavily biassed towards inductively coupled emission and mass spectrometries. It is likely that these routine analytical systems were used in laboratories that did not report the instrumental technique applied as well. Depending on the particular element investigated, sector-field and collision cell inductively coupled plasma mass spectrometry (ICPMS) was used in 10% to >15% of the analytical determinations; about 15% of the mercury determinations were based on cold vapour generation and atomic absorption or fluorescence detection.

Data outputs

The data for each element in the exercise were provided to the participants in tabular and graphical formats for the unknown material, QC04LH4. Data tables consisting of laboratory means, number of observations, associated summary statistics, assigned weights and estimates of within-laboratory variances from the ML consensus mean algorithm, and *z*- and *p*-scores as a function of laboratory were generated, along with the assigned consensus mean and associated expanded uncertainty for each element measured in QC04LH4. These tabular data were complemented by four plots; a raw data plot, a consensus mean plot, a plot of laboratory performance in the *z*- and *p*-score space and a Youden diagram. Example tabular data and corresponding consensus mean plot outputs for Zn in QC04LH4 are given in Table 4 and Fig. 3, respectively. The ML weight is the iteratively derived weight assigned to each laboratory value that comprises the consensus mean. The individual laboratory weights assigned ranged from 65.4% (Laboratory 16) to 99.7% (Laboratory 27) for the Zn example presented. The "Tau Estimate" is an estimate of the within-laboratory variance of the mean that is used in Eqs. 1–3. The assigned *p*-scores are necessarily inversely correlated with the ML weights assigned to each laboratory because of the nature of the consensus mean estimation model employed.

The Zn consensus mean plot (Fig. 3) displays the individual laboratory means and the assigned ML consensus value and its associated expanded uncertainty. This important plot allows laboratories to compare their reported mean value against an estimated congruence value and lends some visual perspective to the heteroscedasticity of the data. Example Mn data plotted in the *z*- and *p*-score space (Fig. 4) shows that this plot serves as an indicator of relative congruence and analytical method repeatability. Laboratories that possess the lowest *p*-scores will be nearest to the abscissa and lab-

Table 4 Consensus mean and associated lower and upper 95% confidence level (CL) range, and laboratory summary statistics, maximum likelihood (ML) weights and *z*- and *p*-scores for Zn in the unknown sample, QC04LH4

Element	Consensus mean	Lower 95% CL	Upper 95% CL
Zn	31.2	30.6	31.8

Lab. #	*N*	Laboratory mean	Standard deviation	ML weight	Tau estimate	*z*-score	*p*-score
1	5	29.4	0.8	0.952	1.13E-01	–0.56	0.25
2	5	31.1	0.6	0.965	8.09E-02	–0.04	0.21
3	10	31.3	2.1	0.841	4.26E-01	0.04	0.67
4	10	30.3	1.5	0.909	2.26E-01	–0.28	0.50
6	5	33.1	0.3	0.991	1.96E-02	0.60	0.09
7	5	34.3	0.4	0.986	3.30E-02	1.00	0.12
8	5	31.8	0.4	0.987	3.05E-02	0.19	0.12
9	5	30.6	0.4	0.988	2.69E-02	–0.20	0.12
10	5	33.6	0.8	0.951	1.15E-01	0.78	0.22
13	5	28.6	1.8	0.767	6.86E-01	–0.83	0.62
14	5	29.5	1.4	0.855	3.84E-01	–0.55	0.47
15	5	30.8	0.4	0.987	2.91E-02	–0.11	0.12
16	5	28.3	2.3	0.654	1.19E+00	–0.92	0.81
17	5	31.2	1.3	0.873	3.30E-01	0.01	0.42
19	5	31.9	1.0	0.926	1.82E-01	0.23	0.30
20	5	30.1	1.0	0.917	2.06E-01	–0.34	0.34
21	5	34.0	0.5	0.980	4.55E-02	0.90	0.14
22	5	29.6	0.4	0.986	3.30E-02	–0.51	0.14
23	8	29.7	0.9	0.954	1.10E-01	–0.47	0.32
24	6	31.4	0.6	0.970	6.97E-02	0.07	0.21
25	5	28.7	1.2	0.877	3.16E-01	–0.79	0.43
27	5	31.4	0.2	0.997	7.91E-03	0.07	0.06
28	5	31.7	0.4	0.983	3.94E-02	0.16	0.14
29	5	33.0	2.4	0.664	1.14E+00	0.59	0.72
31	5	32.7	0.7	0.963	8.70E-02	0.49	0.20
33	5	31.1	0.8	0.952	1.14E-01	–0.04	0.24
Outliers							
11	2	21.800	0	–	–	–3.01	0
12	5	2.702	0.022	–	–	–9.13	0.08
30	1	30.150	–	–	–	–0.33	–
32	1	27.465	–	–	–	–1.19	–

Means and standard deviations are expressed in units of mass fraction, mg/kg

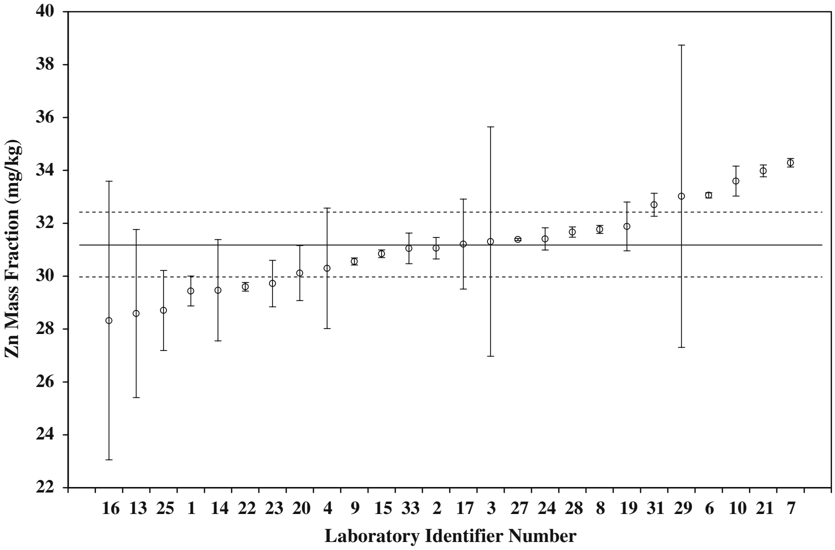

Fig. 3 Consensus mean plot for Zn mass fraction (mg/kg) in the unknown sample, QC04LH4. Laboratory means (*circles*), laboratory standard deviations, consensus mean (*solid line*) and associated expanded uncertainty (*dashed lines*) are shown

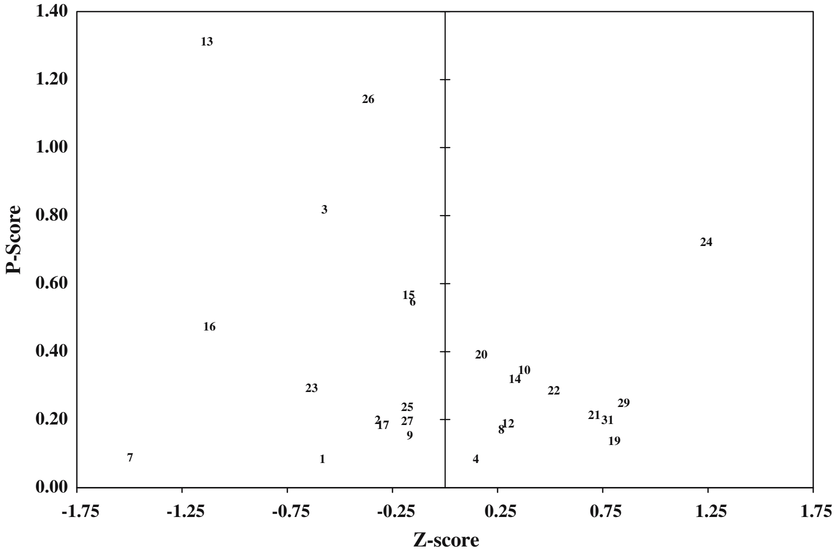

Fig. 4 Laboratory performance scores for Mn in the unknown sample, QC04LH4, plotted in the *z*-score and *p*-score space

oratories that possess the lowest *z*-scores will be nearest to the ordinate axis. The ''floor'' of the abscissa gives a rough indication of the maximum potential heterogeneity for each element in the sample. The Youden diagram (described above) helps laboratories to identify possible systematic and random biasses associated with the analytical method employed. Most trace element data fell in the upper right and lower left quadrants of the Youden diagrams, indicating systematic among-method and/or within-laboratory bias. Mercury and Se measurements were relatively dominated by indeterminate (random) errors.

Data evaluation

The consensus mean algorithm used in this study is designed to minimise the uncertainty surrounding the consensus value through use of an iterative weighting process. Theoretically, this should maximise the utility of comparisons against the consensus mean for each participating laboratory, as the uncertainty cast about the consensus value serves as an indicator of the quality or efficiency of the estimate. If the uncertainty interval is large, the consensus mean estimate becomes inherently less useful from the perspective of the participating laboratories. The tangible effect is that a laboratory would be less likely to alter an in-house method or protocol if there is disagreement between the laboratory and the consensus value when the consensus value possesses a large uncertainty. Consensus data generated using the ML approach were compared to data generated using the robust procedure described previously and Table 5 summarises the data from the comparison. The ML and median consensus estimates

Table 5 Consensus mean and expanded uncertainty (U) data comparison using maximum likelihood (ML) and robust median (Med.) statistics

Element	ML consensus mean	ML relative 95%$U_{k=2}$ (%)	Consensus median	Robust med. relative 95%U^{a} (%)	Consensus estimate ratio (ML/Med.)
Co	0.011	8.34	0.011	31.68	0.998
Cs	0.029	4.75	0.028	11.47	1.044
V	0.047	9.34	0.045	44.19	1.055
Sn	0.059	9.68	0.061	23.21	0.974
Cd	0.221	3.76	0.216	19.80	1.025
As	0.275	7.08	0.288	31.03	0.954
Mo	0.388	2.68	0.390	15.79	0.995
Ag	0.468	4.87	0.474	11.90	0.988
Rb	1.19	2.51	1.18	7.39	1.007
Mn	3.13	2.45	3.08	13.78	1.016
Se	3.37	7.87	3.25	24.65	1.037
Hg	3.60	3.12	3.60	15.94	1.000
Cu	5.20	2.68	5.24	18.15	0.991
Zn	31.18	1.93	31.06	10.60	1.004
Fe	356.6	2.35	353.7	12.34	1.008

Units are expressed as mass fraction, mg/kg

[a] Expanded uncertainty, U, calculated as *k*(MADe), with *k*=1.96

are in very good agreement, as evidenced by consensus estimate ratio values near unity for all of the elements tested. Where the estimates differ significantly is in their traceability and assigned uncertainties. From an efficiency standpoint (after outliers have been removed), it is consistent that the robust MADe approach produces larger confidence intervals relative to the ML approach. It appears that the exercise design and statistical approaches used are at least of similar utility as robust statistics estimates for the data encountered in this interlaboratory comparison. The tighter uncertainty intervals cast about the ML consensus value appear to be reasonable and may be of some use to the participants and the coordinators. For participants, *z*-scores (computed as relative deviations from congruence) become more useful, and if the material is a candidate-certified reference material or other type of quality control sample destined for value assignment, the tighter uncertainty interval assigned to data originating from an interlaboratory comparison may be useful for method development and when combining data from in-house methods and other sources—a common practice in chemical metrology.

The exercise showed that numerous subgroups of the exercise participants demonstrated comparability within the |0 to 1| *z*-range for many elements, based on the use of 10% of the consensus mean as the fixed performance criterion. For any given element, the *z*-score range |*z*|=0 to 1 implies that a laboratory in this subgroup can distinguish between two samples when their respective analyte concentrations differ by 0% to 20%. The *z*-scores are scalable, so any laboratory may wish to challenge their performance using the qualitative IUPAC guidelines. For example, a laboratory that scores z=–0.7 based on a σ_{Target} of 10% of the consensus mean would score z=–1.4 if the performance criterion was tightened to σ_{Target}=5% of the consensus mean. The scaled result in this theoretical example would still be classified as "satisfactory" (|*z*|≤2). The laboratory *p*-scores were typically <10% relative standard deviation for all elements. This type of precision (or better) should be expected for atomic spectrometry measurements. This implies that QC04LH4 is a relatively homogeneous material, as inflated, widely ranging *p*-scores for small or large subsets of laboratories could be indicative of a within-jar homogeneity problem for any particular element.

Group metrics can be used to provide a qualitative mark of performance for the collection of participating laboratories. The average absolute *z*-score [18] and the *z*-score variability, Var(*z*), an analogue to the standard deviation statistic wherein the residuals are replaced with the individual z_i [15], have been proposed as group metrics for the evaluation of interlaboratory comparison data. A view of the data collectively as a function of element reveals some interesting observations. Figure 5 charts several *z*-score and *p*-score group metrics as a function of element (outlier data excluded). The average or median absolute *z*-score is a metric representing the collective congruence of the group of participating laboratories that contributed data to the consensus mean. The highest *z*-scores were observed for As, Se, Sn, V and Co, indicating that these particular elements were the most challenging to measure from a congruence perspective. The Var(*z*) metric is a measure of the width of the *z*-score distribution and, thus, is sensitive to the presence of near outliers. For Se and Co, single measurements (different laboratories) with HS>3 are responsible for the inflated Var(*z*) values. The relatively higher *z* metrics for the

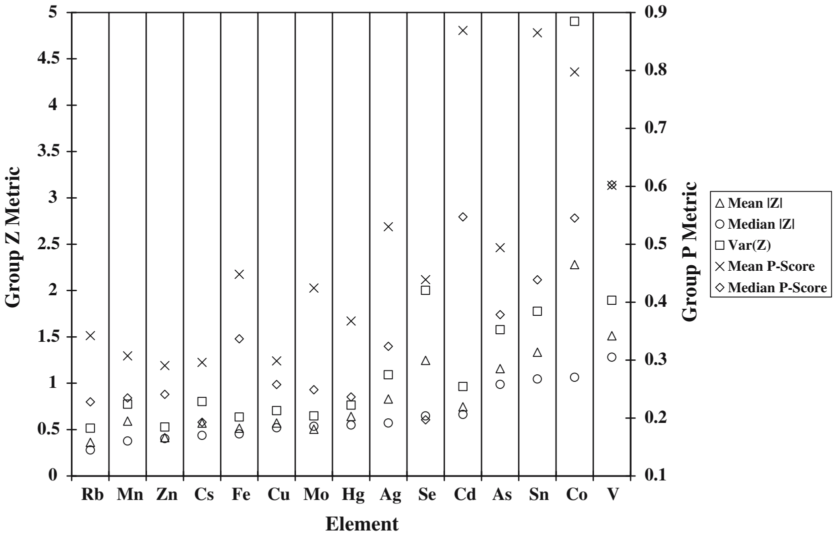

Fig. 5 Group *z*-score and *p*-score metrics for individual trace elements

aforementioned elements could be due to several factors, including the element concentration levels in the sample, random analytical biasses and systematic laboratory biasses, including calibration and blank errors, element volatility or loss, recovery and certain types of chemical and spectral interferences. These factors are influenced by decisions made by the participating laboratories, including the selection of laboratory methods, calibration protocols and instrumentation. It is difficult to specifically pinpoint the analytical problems associated with these elements. If one assumes that ICPMS methods are driving the majority of determinations for these higher average *z*-score elements (see Table 3), and that sample concentrations are not near detection limits (no correlations between any of the group *z*-score metrics and consensus concentrations were observed), it might be possible to deduce that most laboratories view this subset of elements as problematic, with analytical determinations being affected by various matrix or spectral interferences and plasma processes. In addition to suffering from isobaric interferences, As and Se ICP-OES and ICPMS measurements are also subject to differential ionisation effects if the carbon content of the calibration samples and analytical samples is not matched.

Figure 5 also charts the average and median *p*-score as a function of element. The average or median *p*-score is a metric representing the collective precision score of the group of participating laboratories that contributed data to the consensus mean. The highest *p*-scores were observed for Cd, Co, V, Sn and As, indicating that these particular elements were the most challenging to measure from a precision perspective, with ‘‘precision’’ being a proxy for either decreased laboratory and/or method repeatability, or potential sample heterogeneity. It is interesting, but not surprising, to note that several of the high *p*-score elements suffer from large average absolute *z*-scores as well—a trend likely to be related to the tendency for imprecision to track relative bias for small datasets.

Conclusions

The maximum likelihood (ML) consensus mean algorithm and other statistical procedures outlined have been used effectively to evaluate data from the National Institute of Standards and Technology/National Oceanicand Atmospheric Administration (NIST/NOAA) Trace Elements in Marine Mammals Interlaboratory Comparison Exercise. The ML consensus mean processing algorithm and alternative consensus mean calculators are built into the consensus mean command in NIST’s Dataplot statistical software, which is freely available for download on the Internet [19]. The benefits of the ML procedure are that it serves to minimise the contribution to the consensus mean from measurements with poor precision. The interlaboratory comparison scheme described produced consensus data that was in good agreement with data produced using robust median statistics. The ML algorithm, when used in concert with performance assessment tools, such as *z*- and *p*-scores, services the participants by establishing quality benchmark data for the assessment of laboratory performance. These characteristics also benefit reference materials producers by providing good estimates of constituent analyte concentrations and their uncertainties.

The first iteration of this quality assurance exercise in 2000 was a modest endeavour, as only seven laboratories participated. However, recent exercises have demonstrated that the scope of this quality assurance exercise is expanding beyond the interests of the marine mammal contaminants community to the analytical chemistry community as a whole, as numerous domestic and international health, environmental and diagnostic laboratories have been brought in as participants. It is hoped that a core group of these laboratories will regularly participate in future exercises to help underpin and improve the quality of trace element measurements in environmentally important marine biological tissues. International participation has been strong for both the 2003 and 2005 efforts, and it is hoped that this trend will continue. The NIST is currently working on collecting and producing a marine mammal whole blood quality control material for use in future trace element exercises to reflect the increased activity in live animal testing by the marine mammal community. This material will serve to complement sampling efforts focussed on the marine specimen banking of samples, while serving as a challenging quality control matrix that is fit-for-purpose and analytically relevant from both the marine animal health and human health measurements perspectives.

Acknowledgements The following individuals are gratefully acknowledged for their contributions to the Interlaboratory Comparison Exercise. Wayne McFee [(NOAA) National Ocean Service Center for Coastal Environmental Health and Biomolecular Research, Charleston, SC] collected and donated the pygmy sperm whale liver to the NIST through the vehicle of the National Marine Mammal Tissue Bank. Personnel from NOAA National Marine Fisheries Service and the South Carolina Department of Natural Resources assisted with the field sampling effort. This material was subsequently processed into the QC03LH3 quality control material. The white-sided dolphin liver that was processed into QC04LH4 was donated through a similar mechanism by personnel at the New England Aquarium, including Frederick Wenzel. Alan Heckert and Stefan Leigh (NIST Statistical Engineering Division) are gratefully acknowledged for their assistance with the NIST Dataplot software used to establish consensus mean data and plots. David Duewer (NIST Chemical Science and Technology Laboratory) is appreciatively acknowledged for providing scientific guidance on the determination of outliers and the application of weighted and robust statistics procedures to interlaboratory data. Finally, the contributing laboratories are acknowledged for their interest and participation in this interlaboratory comparison exercise program.

Disclaimer: Certain commercial equipment, software or instruments are identified in this paper to specify adequately the experimental procedures. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor does it imply that the referenced items are the best available for the purpose.

References

1. Becker PR, Wise SA, Thorsteinson L, Koster BJ, Rowles TK (1997) Chemosphere 34:1889–1906
2. Zeisler R, Langland JJ, Harrison JK (1983) Anal Chem 60:2760–2765
3. Rukhin AL, Vangel MG (1998) J Am Stat Assoc 93(441):303–308
4. Thompson M, Ellision SLR, Wood R (2006) Pure Appl Chem 78(1):145–196
5. Youden WJ (1959) Ind Qual Control 15:24–28
6. Wijnstekers WA (2005) Reference to the Convention on International Trade in Endangered Species of Wild Fauna and Flora, 8th edn. CITES Secretariat, Geneva, Switzerland
7. International Organization for Standardization (1993) Guide to the expression of uncertainty in measurement, 1st edn. ISO, Geneva, Switzerland
8. Wise SA, Schantz MM, Koster BJ, Demiralp R, Mackey EA, Greenberg RR, Burow M, Ostapczuk P, Lillestolen TL (1993) Fresenius J Anal Chem 345:270–277
9. Greenberg RR, Fleming RF, Zeisler R (1984) Environ Int 10(2):129–136
10. Levenson MS, Banks DL, Eberhardt KR, Gill LM, Guthrie WF, Liu HK, Vangel MG, Yen JH, Zhang NF (2000) NIST J Res 105(4):571–579
11. Mandel J, Paule RC (1970) Anal Chem 42:1194–1197
12. Paule RC, Mandel J (1982) J Res Natl Bureau Stand 87:377–385
13. Schiller SB, Eberhardt KR (1991) Spectrochimica Acta 46B(12):1607–1613
14. Taylor BN, Kuyatt CE (1994) Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST technical note 1297, United States Department of Commerce. Available online at http://www.physics.nist.gov/Document/tn1297.pdf
15. Bednarova M, Aregbe Y, Harper C, Taylor P (2006) Accred Qual Assur 10:617–626
16. Analytical Methods Committee (1989) Analyst 114:1693–1697
17. Analytical Methods Committee (1989) Analyst 114:1699–1702
18. Linsinger TPJ, Kandler W, Krska R, Grasserbauer M (1998) Accred Qual Assur 3:322–327
19. NIST Dataplot. Home page at http://www.itl.nist.gov/div898/software/dataplot/