# Pooled calibrations and retainment of outliers improve chemical analysis

**Andersen, Jens; Alfaloje, Haedar S.H.**

*Published in:*
Proceedings of 1st International Conference on Analytical Chemistry

*Publication date:*
2012

# POOLED CALIBRATIONS AND RETAINMENT OF OUTLIERS IMPROVE CHEMICAL ANALYSIS

Jens E.T. ANDERSEN[a]* and Haedar S.H. ALFALOJE[b]

[a] Technical University of Denmark, Department of Chemistry, Kemitorvet Building 207, DK-2800 Kgs. Lyngby, Denmark.

[b] Department for Environmental Science, Aarhus University, Frederiksborgvej 399, PO Box 358, DK-4000 Roskilde, Denmark

*Received December 20, 2012*

Analytical chemistry has a large responsibility in society, and credibility and reliability are important concepts associated with chemical analysis. Metrology and Quality Assurance (QA) are key areas of interest in contemporary research. Quality in measurements is illustrated by a series of experiments with several analytical technologies comprising of ICP-MS, GC-MS and AAS. The scientific methodology relies on the concept of reproducibility that depends on type of analyte and type of apparatus. By applying the principle of pooled calibrations it is shown that the performance of the apparatus in terms of levels of uncertainty can be tested in a single laboratory. The uncertainties are compared to predictions of the Horwitz formula. It is suggested that this method is universally applicable not only to the actual technologies but also to other technologies in other fields of science. The results indicate that the procedures outlined in the Eurachem/CITAC Guide are of tremendous value to analytical sciences because they direct researcher's attention towards the concept of consensus values rather than towards true values. Introduction of certified reference materials (CRM's) in metrology has provided much new information on working habits in professional laboratories and CRM's may be applied to establish the true level of uncertainty for a given type of analytical method. Finally, it is proposed to devise a new procedure of method validation that facilitates QA in general, thus saving many resources at laboratories.

* Corresponding author: jeta@kemi.dtu.dk

## INTRODUCTION

There is a current scientific interest in developing new methods of quality assurance in analytical chemistry with the aim of creating methods of high degree of reliability. It is now a well-established fact that professional analytical laboratories cannot agree about their results of certified-reference materials (CRM's), which previously been proved by, *e.g.* the IMEP programme of the European Commission.[1-4] An investigation as to the origin of this problem has led the European Commission, among others,[5] to initiate programs of interlaboratory comparisons[6] and metrology.[7] If uncertainty of measurement were underestimated then it would be difficult to agree about results. This is the case when two independent laboratories compares results in terms of confidence ranges. Confidence ranges narrow as a function of number of repetitions, which thus increases the risk of disagreement. This is an analytical paradox; more repetitions provide more disagreement. A careful method validation including uncertainty budgets and checking traceability may overcome this problem when expanded uncertainties (EU's) are developed for comparisons.[8-9] Under the assumption that the method validation included all major contributions to uncertainty ensures EU that any laboratory will be able to reproduce results with a probability of 95 %. This is what is desired for in the development of new analytical methods. Statistical methods are essential to comparisons and comparisons has a need for reliable standard deviation (STDEV) to e.g. t-testing. The reliability depends on the number of repetitions but it also depends on policy with respect to handling of outliers. If outliers were rejected then would the value of STDEV decrease and the validity of comparisons put in jeopardy. Therefore, it cannot be recommended to reject outliers under any circumstances but they should be retained and the effect of their presence on the results should be diminished by adding to the data set more repetitions. The influence of outliers is diminished in this manner and the validity of STDEV is maximized. However, outliers were frequently removed from data sets of interlaboratory comparisons[5, 10] even when the number of repetitions is relatively low.[5] Rejection of outliers is also recommended by ISO 5725 where various statistical means of outlier detection are provided.[10-11] It has been shown that rejection of outliers may produce results that cannot be understood within the frame of investigation; rejection of outliers was supposed to improve quality of measurements but it leads evidently to the opposite. Application of the Hampel test may thus lead to detection of a large fraction, *e.g.* 15 – 55 % of the total number of measurements were outliers that were rejected from the data set,[1] which is generally not reasonable. However show Bednarova *et al.*[1] duly all data both those with and without outliers; an example to follow. Rejection of outliers poses a risk to the credibility of certified reference materials because costumers cannot be expected to reproduce results unless the same number of outliers was rejected in independent series of measurements. Other investigations[2-4] mentions never outliers thus implying they were neither observed nor rejected from the data set. Instead, results of professional laboratories were divided into categories of

satisfactory, questionable or unsatisfactory as derived by the aid of z-scores[3], Grubb's test[10] or Hampel test.[1] Again, a large fraction of results, more than a third, of participating laboratories fell into the groups of either questionable or unsatisfactory,[3] which may seem incomprehensible or strange at least to rejected laboratories.

Recent investigations show that another pathway could be followed by not rejecting outliers.[12-15] It would be feasible in a manuscript to consider both cases, and it is also worthwhile to consider a 'worst-case scenario' where all data of several series of independent measurements were included in the analysis. The reason for evaluating of such a scenario is a result of recognizing relatively large uncertainties of measurements performed over long periods of time. If the calibration was working properly in all experiments then would instrumentally induced fluctuations be cancelled out but this seems not to be the case for all types of apparatuses. Therefore a series of experiments were conducted with the aim of evaluating the validity of pooled calibrations.[12-15, 16] Mostly is quality assurance is reserved for specialized journals but it should be included in every single publication with method development as a means of increasing credibility and reliability. Initial results indicate full correspondence between predicted uncertainties and those obtained from multiple repetitions, which may be a demonstration of statistical control. Results of ion chromatography[15] indicate that relative uncertainties comply with the general coefficient of variation predicted by Horwitz' formula.[17]

## THEORY

Almost every analysis of contemporary analytical chemistry is performed by using a straight line for the operational calibration of the apparatus. In this way, the response of the detector is depicted as a function of concentration within a certain range of responses and concentrations given by a limit of detection and a maximum concentration. The limit of detection may be regarded as a limit that is unattainable for routine analysis in the sense that determination of concentrations of a solution of unknown concentration, henceforth briefly denoted as the unknown, provides results of low accuracy (large STDEV's). Although precision might be good of measurements performed by using concentration close to the LOD, then are the accuracy of results frequently unacceptable for decision making. Therefore, there is a need to establish a new concept that defines the minimum concentration where the accuracy is acceptable and, correspondingly, it would also be interesting to establish a rigid value for the maximum concentration of analysis. The minimum and maximum value of analysis may be denoted as lower limit of analysis (LLA) and upper limit of analysis (ULA), respectively. In order to establish such concepts, it is worthwhile to consider the real detector response, which never is linear. All detector responses are non-linear and, for the particular case of absorbance measurements, the response (y) may conveniently be described by the equation:[13]

$$y = A \cdot (1 - \exp[B \cdot (x - x_0)]) \qquad B < 0 \tag{1}$$

where A is the maximum response of the detector, B is the characteristic concentration, x is the concentration and $x_0$ is the intercept value. The tangential equation that is valid at concentrations below $x_0 - \frac{1}{B}$ may be obtained by expanding eq. 1 to first order which yields:

$$y \cong -A \cdot B \cdot x + A \cdot B \cdot x_0 \quad \text{at} \quad 0 < x \leq x_0 - \frac{1}{B} \tag{2}$$

However, eq. 2 cannot be used as a calibration line as illustrated in Fig. 1 showing the tangential line that approximates the response curve of eq. 1 together with the regression line. The regression line is depicted in the same range of concentrations as that of the tangential line, which indicates that it may be used as a calibration line. In fact may all sorts of calibration lines be constructed from linear approximations to eq. 1 but then arises a few problems that must be dealt with before deciding on the range of concentration for operational calibration. In theory deviates the regression line of Fig. 1 significantly from the response curve with exception of the two points of intersection. Accordingly, these systematic deviations must be smaller than the uncertainty given by fluctuation of the experimental data, which may be true for pooled calibrations.[12-15] In general, the range of calibrations may be chosen, as to certify, that these systematic deviations were smaller than the experimental spread of data but this may impose restrictions on the calibration which result in a very narrow range of concentrations.

Addition of experimental data of multiple calibrations provides a graph with a spread of data around the response curve which reflects the genuine performance of the apparatus (Fig. 2a).[13] A single calibration line seems to be inadequate for describing the level of uncertainty involved in the analysis but the application of pooled calibrations may create a satisfactory correspondence between predicted uncertainties and those observed by experiment.[12-15] In Fig. 2a is the response curve shown together with a tentative spread of experimental data. The regression line (Fig. 2b) is then constructed within the range of concentrations given in eq.2. For the sake of simplicity is the STDEV associated with the data of Fig. 2b modeled by a straight line with slope ‘a’ and intercept ‘b’ (Fig. 2c).

Manufacturers apply non-linear regression to the calibration curve of Fig. 2c only for a few types of apparatuses such as AAS. Researchers use most frequently linear regression but the method, by which the calibration line was determined, is rarely described in much detail. However, assessment of the full range of calibrations and the associated uncertainty of measurement must be evaluated in detail during the method validation. A straightforward method is therefore proposed and it relies on functions of MSExcel spreadsheets. An overview of the calibration curve is produced by using many standards and also standards of very high concentration in order to recognize the actual course of the regression curve (Fig. 2a). The ULA

may be found by fitting the curve of eq. 1 to the experimental data (Fig. 2a) by the aid of MSSolver; this gives the ULA according to eq. 2. The LOQ is supposed to be the ultimate lower limit of analysis but it was recently suggested that it is not convenient for analysis because the uncertainty of analysis might be too large. [12-15] The core of the problem of calibrations is related to the concept of uncertainty. In the following is the standard uncertainty represented by the STDEV of concentration derived upon the basis of pooled regression lines. According to GUM[8] and to the Eurachem/CITAC guide,[9] is the expanded uncertainty then equal to the STDEV multiplied by a factor of two.

Tentatively, a straight line may be fitted directly (not shown) to the data of Fig. 1a and, dependent on the purpose of the analysis, this regression line may be used for calibrations. However, this would result in very large uncertainty on measurements and a risk of introducing bias. In addition, would such a regression line exhibit an intercept significantly different from the value of the blank. Accordingly, remains the task to find the regression line, that produces the lower level of uncertainty. At first glance the data of Fig. 1a indicates that the lower level of uncertainty was obtained at low concentrations and this is also the range of concentrations where absolute-standard deviations are small. However, the absolute uncertainty is not of much interest when it is the purpose estimating the best possible concentration range to calibrations. It is the relative uncertainty that must be estimated, which is not necessarily found within the range of low concentrations. Thus is a more careful analysis needed to evaluate the relative uncertainty as a function of concentration.

Empirically, within the linear range of responses, the STDEV ($s_y$) is proportional to concentration (Fig. 2c). Another empirical formula of non-linear relation between concentration and STDEV is provided by the Eurachem/CITAC[9] and this may also be used but it is slightly more laborious to use for the analysis. The slope and intercept of the regression line of Fig. 2c is given by a and b, respectively, and the STDEV of response values ($s_y$) becomes:

$$s_y = a \cdot x + b \tag{3}$$

The STDEV of responses (eq. 3) constitutes part of the RSD of concentration ($s_x/x$), which becomes:[12]

$$\frac{s_x}{x} \cdot 100 = \frac{100}{\alpha \cdot x} \cdot \sqrt{\frac{s_\beta^2 + (a \cdot x + b)^2 + x^2 \cdot s_\alpha^2}{2}}\ \% \tag{4}$$

The factor of two in denominator of equation under the square root is introduced in order to average the contributions to the STDEV.[14, 18] Correlation terms were also accounted for by averaging, and correlation terms[18] were thus omitted in eq. 4. At concentrations higher than b/a (x > b/a), which is most

frequently the case, then $s_x/x \cdot 100$ itself approaches a constant value that is denoted as the minimum RSD or best RSD (BRSD in percentage):

$$\text{BRSD} \rightarrow \frac{100}{\alpha} \cdot \sqrt{\frac{a^2+s_\alpha^2}{2}}\ \% \qquad \text{when } x > \text{SBR (eq. 6)} \tag{5}$$

The constant value of eq. 5 is obtained when the influence of $s_\beta$ on $s_x/x$ becomes negligible, which occurs at the concentration denoted as the start-of-best range (SBR):

$$\text{SBR} \cong \frac{s_\beta}{\sqrt{a^2+s_\alpha^2}} \tag{6}$$

At the concentration of the SBR (eq. 6), the RSD is then given by:

$$\text{RSD(SBR)} = \frac{s_{SBR}}{SBR} = \frac{100}{\alpha} \cdot \sqrt{a^2 + s_\alpha^2}\ \% \tag{7}$$

Which is some 40 % higher than the BRSD but in the present context is this deviation considered as a minor difference because differences in RSD's of interlaboratory testings may amount to much higher values.[19] Therefore, seems the suggestion of a constant RSD above the SBR valid. It should be noted that all the equations derived above are only valid for pooled calibrations where a thorough method validation has provided the genuine spread of data over long-periods of time with multiple independent calibrations. Within the range of concentrations defined by SBR < x < ULA, is the best concentration range established for linear calibrations. It provides the lowest possible RSD that is claimed to be a characteristic method parameter that cannot be superseded by choosing another range of calibrations.

It is common practice to require that the intercept of the calibration line should correspond to the response value of the blank. That is, the intercept (β) not deviate significantly from blank values, which may be expressed by the condition:

$$\bar{y}_0 - 2 \cdot s_{\bar{y}_0} < \beta < \bar{y}_0 + 2 \cdot s_{\bar{y}_0} \tag{8}$$

where $\bar{y}_0$ and $s_{\bar{y}_0}$ represent the average value and standard deviation, respectively, of multiple independent determinations of blanks. Measurements used for determination of LOQ cannot be used to calculate parameters of eq. 8 because blanks are measured consecutively within the same series of measurements in

order to calculate the LOQ whereas independent series of measurements are used for eq. 8. Negative blank values may occasionally be obtained and they should also be included in the series of measurements used for determination of parameters of eq. 8. Ideally, if the intercept (β) were zero in the type of experiments, that have a response curve of the type shown in Fig. 1a, then would an average-response value ($\bar{y}_0$) be slightly positive with both positive and negative values found in the series of 10 measurements.

Now it remains to determine the approximate value of uncertainty of the unknowns that may be estimated by eq. 4. It should also be noted that the concentration of unknown cannot be determined on the basis of the regression line of pooled regressions. The regression line of pooled calibrations is calculated by MSExcel spreadsheet using the dataanalysis toolpack/regression that provides $\alpha$, $s_\alpha$ and $s_\beta$ for eq. 4. Thus, is used the average value of concentrations of unknowns determined in several independent series of measurements to inserting for 'x' in eq. 4, in order to estimate RSD of c, that is to estimate $\frac{s_x}{x} \cdot 100$ of the unknown. The RSD(cal) of unknown, determined by pooled calibrations, should then correspond to the RSD(rep) of unknowns, that were determined in multiple independent series of measurements where the concentration of unknown of each experiment was determined by the aid of a single calibration line.

## EXPERIMENTAL

GC – MS

*Chemicals*

Analytes were dissolved in ethyl acetate ($C_4H_8O$, CAS 141-78-6, Sigma Aldrich) and three analytes comprising of dimethyl sulfoxide ($C_2H_6OS$, CAS 67-68-5, Sigma Aldrich), butanoic acid ($C_4H_8O_2$, CAS 107-92-6, Fluka), and 3-heptanol ($C_7H_{16}O$, CAS 589-82-2, Fluka) were applied to quantitation.

*Apparatus*

The measurands were composed of the analyte dissolved in ethyl acetate and they were measured by a Finnigan GCQ Polaris (Thermo-Finnigan Coorporation, San Jose, California) equipped with ion-trap detector with a detector voltage of 1.9 kV and background pressure of $10^{-5}$ torr. The ion-source temperature was 200°C and the ionization mode was + EI at 70 V. A full scan mode was used for the mass spectrometer using range of m/z 35 – 300. The sample volume was 1 μL in the A200S autosampler where

injections were performed with 0 air volume, 3 pre-injection wastes, 2 sample washes, pre-injection time 3 and post-injection time 0. Chromatographic separation was obtained by a GsBP20M-carbowax column of dimensions 30 m x 0.25 mm x 0.25 microns. Helium of 99.995 % purity was used as carrier gas at flow rate 14 mL/min. Injection was performed using a split ratio of 1:16 with a 4 mm glass liner with a plug of glass wool. The temperature of injection was 300°C and the temperature of transfer line was 275°C. The oven was programmed to start at 40°C for 5 min. and linearly increase temperature to 80°C at 5°C/min. followed by an increase to 200°C at 15°C/min. and finally increase temperature to 280°C at 10°C/min. at a split ratio of 10.

ICP-MS

*Chemicals*

A stock solution of standards containing elements Na, Mg, P, Ca and Sr was prepared by dilution of P/N 4400 ICP-MS ICS solution A (500 mg/L Al, Ca, Fe, K, Na, Mg, P, S, 1000 mg/L C, 3600 mg/L Cl and 10 mg/L Mo and Ti, CPI International) and Sr PlasmaCAl (1000 mg/L, SCP Science) with Millipore water (18.2 MΩ). The P/N 4400 ICP-MS ICS solution A is traceable to the NIST SRM 3100 series. Multi-element standards were prepared with concentrations 0, 1, 5, 10, 25, 50, 75, 100, 150, 200, 225, 250, 300, 400, 500, 800, 1000, 2000, 3000, 5000 µg/L. Rh was added as internal standard in the form of a solution (1 mg/L) prepared by dilution of Rh-standard solution (1000 mg/L, Fluka, CAS 10139-58-9, Lot. 1139119, filling code 54405280), with a concentration of 10 µg/L Rh within each calibration standard. The solution of unknowns were prepared by dilution with 0.1 M nitric acid (Merck, suprapur, 65 %) of dissolved (Anton-Paar High-pressure microwave oven MW 3000) bone meal (NIST SRM Bone Meal 1486).

*Apparatus*

Concentrations were determined by a SCIEX Elan 6000 (Perkin-Elmer) apparatus equipped with cross-flow nebulizer. The RF power was 1000W and the argon flow was 20 L/min. and 0.95 -1.00 to the nebulizer. The background pressure was $1.3 – 1.8 \cdot 10^{-7}$ Torr and the working pressure was $1.4 – 1.6 \cdot 10^{-5}$ Torr. Mains-water temperature was 16-19°C and the torch-box temperature was 40-43°C. Data were recorded by the following settings in the quadrupole-peak-hopping mode: MCA channels/peak = 1, number of replicates = 3, dwell time = 100 ms, readings/replicate = 10, dual detector modem, resolution 0.5 - 0.7 and lens voltage 7.5-8.75 V. The following isotopes were measured by ICP-MS: $^{24}$Mg, $^{25}$Mg, $^{26}$Mg, $^{31}$P, $^{42}$Ca, $^{43}$Ca, $^{88}$Sr and $^{103}$Rh.

## RESULTS AND DISCUSSION

The uncertainty of a result obtained by the aid of a single calibration line may be denoted as the uncertainty related to short-term precision whereas the uncertainty of long-term precision and accuracy is the subject of the present investigation.

In the series of GF-AAS measurements published earlier,[14] it was shown that a calibration line determined by the aid of uncertainty estimates could be considered as linear when data of high concentrations were eliminated from the set of data at a step-wise manner. The first step was to construct the regression line within the range of concentrations defined by first order approximation (eq. 2) to the response curve (eq. 1), which yielded the values shown in Table 1. This first step produced a calibration line with a positive intercept that was different from blank values. In step two was eliminated data of high concentrations from the data set, which produced another calibration line a calibration line with a slightly narrower range of calibrations was obtained, but again, the intercept was significantly different from the blank value (Table 1). By imposing the condition of eq. 8 to the series of data representing the response curve,[14] produced elimination of data at high concentrations a calibration line with an intercept equal to blank values but with an even narrower range of concentrations that were suitable for calibrations (Table 1). The step-wise procedure outlined above is straightforward, and the data of high concentrations eliminated during evaluation of the calibration range, does not correspond to removal of outliers. Eliminated data cannot be considered as outliers because uncertainty is not affected upon their removal. Further, the concentration range of calibration may be extended considerably by using the response function of eq. 1 directly at the expense however of more complicated uncertainty estimates.[13]

At the concentration given by the LLA is the REU 100 %[12] and the REU(cal) is approximately constant (approx.12 - 17 %, Table 1) at high concentrations, which is a characteristic feature of many types of apparatuses used for chemical analysis. According to Table 1 it is not possible to analyse at an REU(cal) lower than 12 %, which conflicts with results published earlier.[20] However, a very good correspondence between REU(cal) of the present work and corresponding Horwitz[17] values shows that it is possible to estimate the universal uncertainty of determination of iron by GF-AAS without the aid of an uncertainty budget. (Eurachem/CITAC) The optimum range of calibrations is therefore determined as 30 μ/L (SBR) to 300 μ/L (ULA) where the REU(cal) is 16 % (Table 1). It is essential for reliability and trueness that the concentration is determined of a CRM during method validation. For every single calibration line the concentration is determined of the CRM, which yields a pool of independent measurements of the CRM. The REU of these repetitions may be denoted as REU(rep) and it must correspond to the REU(cal) at least within an order of magnitude. A full method validation includes thus an evaluation of REU(cal) and REU(rep). When REU(cal) is equal to REU(rep) then is said the investigation to be in statistical control. A strategy of all future experiments may therefore be summarized as follows:

1. Perform a thorough method validation to provide the REU versus concentration (Fig. 1).
2. Prepare a few blanks, a single standard, a solution of a CRM and unknowns.
3. Perform measurements.
4. Evaluate data in terms of uncertainties given by the method validation (1).

The absolute STDEV derived from the method validation outlined above is used to perform testings and comparisons, e.g. t-testing, which is essential to the art of decision making. An assessment of bias or an investigation of influence of interferences on results would require differences that exceeded those shown in Table 1.

Analysis of 3-heptanol, dimethyl sulphoxide and butanoic acid was performed by GC-MS and method validation was performed according to the procedure described above (Table 2). A standard was prepared as 'unknown' in a manner similar to the other standards to the GC-MS experiments. The figures of merits were determined by a full method validation and pooled calibrations (Fig. 2b and Table 2). Although only a few repetitions were applied to the analysis it was found that the system was roughly in statistical control (Table 2). The reliability of the REU(rep) was only 59 % owing to the low number of repetition; it cannot be expected to obtain full correspondence between REU(rep) and REU(cal) after only four replicates (Table 2). However, the compliance between BREU, REU(cal) and the REU of the Horwitz[17] formula was excellent, which supports the notion of uncertainty estimates from pooled calibrations also works in the case with GC-MS analysis.

An in-depth method validation was also performed in conjunction with analysis of several isotopes by ICP-MS (Table 3). A very distinct discrepancy was identified between the LOQ of the manufacturer and the LLA and SBR in the present investigation. Since the LLA represent the minimum concentration of analysis where REU is less than 100 % it was surprising to find that the LLA could be greater than LOQ by a factor of up to 2600 (Table 3). An LLA value of 3.7 μg/L for $^{88}$Sr was a factor of 720 higher than the corresponding LOQ value of the manufacturer but it corresponded perfectly well to the LOQ derived by Kruger *et al.*[21] using a sector-field ICP-MS for analysis of trace elements in digested human-follicular fluid. Such large deviations were unexpected and they bing the validity of results that may have been determined at low concentrations in jeopardy. The isotope $^{88}$Sr is known to be only vaguely influenced by the presence of interferences, such as oxides, nitrides carbides and doubly charged ions, which was confirmed by the results of Table 3.[22-27] Other isotopes, in particular those of lighter elements such as Na, Mg and Ca are prone to be influenced by interferences, which were also found during the present investigation (Table 3). Although the isotopes of light elements were expected to be greatly influenced by interferences, the concentrations of unknowns were determined at acceptable accuracy (Table 3). All concentrations except for $^{43}$Ca were recovered but the REU were of considerable magnitudes. The minimum REU was found for the isotope $^{88}$Sr where it was found that these measurements were the only ones in statistical control. The

BREU's of $^{88}$Sr and 31P were calculated as 24 % and 26 %, respectively, in excellent agreement with REU of CRM's (Table 3). Despite the large number of data used to construct calibration lines (N >220) and the large number of repetitions (N = 50) the BREU and REU of CRM differed by up to a factor of 4.4 ($^{26}$Mg/IS). Such a discrepancy is considered to be small in the present context where differences in uncertainties easily mount to order(s) of magnitudes.[19] Addition of correlation term[9, 18] that adjusts for the experimental correlation between slope and intercept caused not substantial changes to REU(calibration) in Tables 2 and 3. EU's calculated by the Horwitz formula were 4 %, 4 %, 3% and 2 % for elements Na, Mg, P and Ca, respectively which were almost an order of magnitude lower than REU (and BREU) of Table 3. It is known that the Horwitz formula may provide too optimistic uncertainties even at the trace level of concentrations[28] but determinations might improve by considering, in more detail, the interferences influence on results, which was not attempted in the present investigation.

An elevated BREU as compared to REU of CRM both with and without IS's is interpreted as a general need to construct a calibration line for each analysis, which effectively eliminates some types of interferences and short-term variations of the apparatus. However it may be suggested that future analysis of Sr by the $^{88}$Sr isotope and analysis of P by $^{31}$P require only a blank, a single standard and a number of repetitions of the unknown. The result thus obtained can then be associated with an uncertainty given by the method validation of Table 3, which decreases cost of analysis for routine analysis. Only one isotope of calcium ($^{42}$Ca) showed an acceptable level of recovery but the results of both isotopes were obtained by measurement at concentrations of twice the ULA values, where reliability is in jeopardy. Dilution of calcium-unknowns provided results of poor precision and accuracy (not shown), which demonstrates that calcium is strongly influenced by spectral interferences.[22-27] Since the number of repetitions was not displayed in the certificate of NIST SRM Bone Meal 1486 it is in principle not possible to make a statistical comparison between results and certified values; therefore can it not be stated whether or not the results obtained differed from certified values (Table 3). The main drawback of analysis of Na, Mg and Ca is the high level of uncertainty which means that many repetitions were required in order to obtain an accurate concentration. It was also shown that analysis including an internal standard (IS) (10 μg/L of Rh) did not improve the analysis in terms of accuracy and level of uncertainty. Since the intensity of parent element is divided by the intensity of the IS in the data treatment, it is expected that the uncertainty increases according to the law-of-propagation of uncertainties.[12] A single result ($^{23}$Na) even showed even an increase in REU from 20 % to 72 % by using of IS in the analysis. The data of Table 3 neither proves nor disproves this general expectation but it should be noted that no major advantage was gained by using an IS. Thus it does not seem worthwhile to introduce IS; it is simply too laborious and the gain with respect to uncertainty is, at best, negligible. The results of Table 3 thus suggest that ICP-MS exhibits good accuracy for some elements (Na, Mg, P and Sr) but Ca was not analysed very well unless measured at very high concentrations above ULA. But the Best REU (BREU) and REU of CRM corresponded to coefficients of variation obtained in

environmental investigations.[20] Phosphorus was also determined at an REU that was comparable to results of CRM's.[29]

The influence of number of repetitions was investigated by experiments of F-AAS where Co was analysed in many independent series of experiments. Two CRM's were applied to the analysis and recovery was very close to 100 % for both EUH-1 waste water and EPH-1 drinking water; concentrations were determined as 0.72 ± 0.05 mg/L(STDEV) and 0.096 ± 0.008 mg/L(STDEV) after 276 and 261 repetitions, respectively. The consensus values were 0.74 ± 0.02 mg/L and 0.095 ± 0.003 mg/L, respectively. It was found that average values changed as a function of number of determinations, where the average value of EUH-1 decreased (Fig. 3a) whilst the average value of EPH-1 exhibited a mostly increasing tendency (Fig. 3b). The results of Figs. 3a and 3b thus indicate that more than 50 repetitions were required to produce reliable average values. In spite of the very large number of repetitions was the value of EUH-1 expected to become closer to the consensus value. The standard deviation of the mean is only 0.003 whereas the observed difference between average value and consensus value is 0.02. Most likely the deviation arises as a result of outlier rejection during manufacturer's assessment of consensus values of the CRM's, according to the certificate of analysis. However, t-testing shows that the final result of EUH-1 differed significantly from the consensus value whereas complete agreement was found for EPH-1. In Figs 3c and 3d is shown the development in REU during multiple determinations of EUH-1 and EPH-1. The REU is shown in chronological order and the initial REU of 0.23 % (EUH-1) and 13 % (EPH-1) after the first five measurements increased to 6.3 % and decreased to 8.6 %, respectively, after approx. 100 repetitions. Similar to average values (Figs. 3a and 3b) is this a demonstration of the law of large numbers in statistics.[30] The only difference is the large random deviations during the first series of experiments. The results of Figs. 3a are important because they show that the average value that was originally in compliance with the consensus value had decreased to a level significantly different from the consensus value and vice versa (Fig. 3b). Therefore it is astonishing to recognize that most investigations of contemporary analytical chemistry introduces only a single calibration line and a low number of repetitions but still obtains a full correspondence between measurements and expected values. The results of the present analysis show that this, most likely, not is possible. This problem may further be illustrated by a result obtained during the analysis of Sr by ICP-MS where an excellent calibration line (not shown), with a regression-coefficient squared $R^2$ of 0.99998, was found for the isotope $^{84}$Sr. The LOQ was determined as 6.5 μg/L and manufacturer's LOQ was approx. 0.5 μg/L in compliance with the corresponding value for $^{88}$Sr. Before multiplying with the dilution factor, the concentration of the unknown was determined as approx. 9 μg/L, which provided a concentration of 263 ± 40 μg/L(STDEV) (at four repetitions, N = 4) for the unknown. With a certified value of 264 ± 7 μg/L were the found and certified values in perfect alignment. However, an LLA of 35 μg/L, SBR = 130 μg/L, and BREU = 20 % of pooled calibrations indicate that further investigations

were required, and a result of 180 ± 140 μg/L was also found after 50 repetitions using between two and four repetitions for each calibration line. Since the confidence interval is 40 μg/L of this latter result, it was significantly different from the certified value, according to conventional statistics. However, the result may be reported as 180 ± 280 μg/L when the EU (coverage factor, k = 2) is considered and now, the result is in agreement with the certified value. This final result (180 ± 280 μg/L) reveals that the level of dilution was not practical because too many repetitions were required in order to obtain a reliable value, owing to the high level of uncertainty. However, the analysis of this particular case shows that it was possible to obtain a result that was apparently correct according to contemporary methods of analysis but it was actually incorrect after a thorough method validation of pooled calibrations. The experiment also demonstrates the uselessness of the regression coefficient for characterization of uncertainty, and it is suggested to avoid reporting this particular figure of merit in future investigations.[31-32]

Operational calibrations are performed with the aim of eliminating systematic differences between results obtained in consecutive series of experiments. It may be stated that calibrations takes care of long-term variations of the apparatus influence on results. This expectation requires that results correlate with slope of calibration line. Otherwise, the efficacy of calibration would be limited and it would not be necessary to perform calibrations in the first place. In Figs. 4a and 4b is shown the correlation diagram of slope and results of unknowns that corresponds to the results obtained in Figs. 4a and 4b. The correlation-diagrams (Figs. 4a and 4b) clearly demonstrate that slope of calibration line and results of unknowns were uncorrelated within certain limits that are defined by the EU of slope and expanded uncertainty of results, as indicated by ellipses (broken lines). Accordingly, produces a steep calibration line not necessarily a low value of the unknown and vice versa. Approximately 4 % of the data were found to reside outside the limits defined by the ellipse of EU's (Figs. 4a and 4b) which corresponds well to the distribution of the conventional 95 % - confidence limit. Since the plot of Figs. 4a and 4b provide a convenient overview of the efficacy of calibration it may proposed that this type of plot is used in future method validations.

Results that were published earlier show the same trend as those presented in Table 1. Acceptable correspondences were found for several different technologies when the method validation was performed according to the principle of pooled calibrations.[12-15] In order to obtain correspondence between calibration-uncertainty (REU(Cal)) and repetition-uncertainty (REU(rep)) it is imperative to retain all outliers in the analysis. Otherwise the uncertainty estimates become unreliable and the system may not be in statistical control. It is therefore recommended to keep all measurements and to never discard potential outliers whatsoever from the data set.[2, 11]

## CONCLUSION

The principle of pooled calibrations was used to perform method-validation of several different analytical technologies comprising of GF-AAS, AAS, GC-MS and ICP-MS. Pooled calibrations were used to predict the uncertainty of measurement in terms of EU's that should be used for comparison of methods and a satisfactory correspondence between predicted and measured uncertainties was found in most cases. The correspondence was less pronounced in cases where low numbers of repetitions were available and in cases where results were influenced by interferences. A best linear range of concentrations for chemical analysis was proposed from the uncertainty analysis of pooled calibrations and a new set of figures of merits was introduced (eqs. 5 – 7). It was thus shown that the total extension of the calibration line could be established on the basis of an evaluation of uncertainties. In order to obtain full correspondence between predicted (REU(cal)) and measured uncertainties (REU(rep)) was it important to retain all outliers and perform many measurements, preferably more than one hundred. Otherwise were both average values and uncertainties incorrectly estimated thus jeopardizing trueness and reliability. The slope-versus-result correlation diagram (Fig. 4) proved to be useful in order to provide an overview of long-term stability of the detector response. It was suggested that long-time variations in detectors provide the major contributions to total uncertainty. It is also proposed that the concept of pooled calibrations and retainment of outliers may radically change scientific methodology in terms of changing focus from short-term precision and accuracy to long-term precision and accuracy.

*Acknowledgements:* The financial support of Director Ib Henriksen's Foundation, P.A. Fisker's Foundation and the Idella Foundation is gratefully acknowledged. Many thanks are due to Anders C. Raffalt, Søren R. Sørensen and Francisca Henriksen for their valuable contributions.

## REFERENCES

1. M. Bednarova, Y. Aregbe, C. Harper and P. D. P. Taylor, *Accred. Qual. Assur.*, **2006**, *10*, 617–626.

2. E. Vassileva and C. R. Quétel, *Anal. Chim. Acta*, **2004**, *519*, 79–86.

3. M.B. de la Calle Guntinas, A. Semeraro, I. Wysocka, F. Cordeiro, C. Quetel, H. Emteborg, J. Charoud-Got and T.P.J. Linsinger, *Food Add. Cont.*, **2011**, *28*, 1534–1546.

4. M.B. de la Calle Guntinas, I. Wysocka, C. Quetel, E. Vassileva, P. Robouch, H. Emteborg, and P. Taylor, *Tr. Anal. Chem*., 2009, 28, 454-465.

5. S. J. Christopher, R. S. Pugh, M. B. Ellisor, E. A. Mackey, R. O. Spatz, B. J. Porter, K. J. Bealer, J. R. Kucklick, T. K. Rowles and P. R. Becker, *Accred. Qual. Assur.*, **2007**, *12*, 175–187.

6. http://irmm.jrc.ec.europa.eu/interlaboratory_comparisons/Pages/index.aspx (accessed October 31, 2012)

7. http://www.emrponline.eu/ (accessed October 31, 2012)

8. BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO and OIML, 2008 *Evaluation of measurement data—guide for the expression of uncertainty in measurement.*

   http://www.bipm.org/en/publications/guides/gum.html (accessed October 31, 2012)

9. A. Williams, S. Ellison, M. Berglund, W. Haesselbarth, K. Hedegaard, R. Kaarls, M. Mansson, M. Rosslein, R. Stephany, A. van der Veen, W. Wegscheider, H. van de Wiel, R. Wood, P.X. Rong, M. Salit, A. Squirrell, K. Yasada, R. Johnson, J.-. Lee, D. Mowrey, P. De Regge, A. Fajgelj and D. Galsworthy, *Eurachem/CITAC Guide*, Quantifying Uncertainty in Analytical Measurement, 2nd edition, 2000. http://www.measurementuncertainty.org/mu/QUAM2000-1.pdf (accessed October 31, 2012)

10. M. Gerboles, D. Buzica, R.J.C. Brown, R.E. Yardley, A. Hanus-Illnar, M. Salfinger, B. Vallant, E. Adriaenssens, N. Claeys, E. Roekens, K. Sega, J. Jurasović, S. Rychlik, E. Rabinak, G. Tanet, R. Passarella, V. Pedroni, V. Karlsson, L. Alleman, U. Pfeffer, D. Gladtke, A. Olschewski, B. O'Leary, M. O'Dwyer, D. Pockeviciute, J. Biel-Ćwikowska, J. Turšič, *Atmospheric Env*., **2011**, *45*, 3488 – 3499.

11. International Standardisation Organisation (ISO) 5725, 1st edition, Geneva, Switzerland, 1994.

12. J.E.T. Andersen, *J. Chem. Ed.*, **2009**, *86*, 733-737.

13. J.E.T. Andersen, *Microchim. Acta*, **2008**, *160*, 89-96.

14. J.E.T. Andersen, *J. Anal. Chem.(Russia)*, **2008**, *63*, 308-319.

15. J.E.T. Andersen, M. Mikolajczak, K. O. Wojtachnio-Zawada and H. V. Nicolajsen, *J. Chrom. B*, **2012**, *908*, 122-127.

16. L. Brüggemann, P. Morgenstern and R. Wennrich, *Accred. Qual. Assur.*, **2005**, *10*, 344-351.

17. W. Horwitz, *Anal. Chem.* **1982**, *54*, 67A-76A.

18. L. Cuadros-Rodríguez, L. Gaèmiz-Gracia and E. Almansa-Loèpez, *Tr. Anal. Chem.*, **2001**, *20*, 195-206.

19. I. Papadakis, J. V. Nørgaard, E. Vendelbo, L. Van Nevel and P. D. P. Taylor, *Analyst*, **2001**, *126*, 228–233.

20. P. S. Sande Fouz, J. M. M. Avalos and E.V. Vázquez, *Comm. Soil Sci. Pl. Anal.*, **2012**, *43*, 280-287.

21. P. C. Kruger, M. S. Bloom, J. G. Arnason, C. D. Palmer, V. Y. Fujimoto and P. J. Parsons, *J. Anal. At. Spectrom.*, **2012**, *27*, 1245-1253.

22. D. De Muynck and F. Vanhaecke, *Spectrochim. Acta Part B*, **2009**, *64*, 408-415.

23. G. De Wannemacker, A. Ronderos, L. Moens, F. Vanhaecke, M. J. C. Bijvelds and Z. I. Kolar, *J.Anal. At. Spectrom.*, **2001**, *16*, 581-586.

24. M. Kovacevic, W. Goessler, N. Mikac and M. Veber, *Anal. Bioanal. Chem.*, **2005**, *383*, 145-151.

25. J. S. Becker, K. Fuellner, U. D. Seeling, G. Fornalczyk and A. J. Kuhn, *Anal. Bioanal. Chem.*, **2008**, *390*, 571-578.

26. A. P. Krushevska, Y. Zhou, V. Ravikumar, Y. Kim and J. Hinrichs, *J. Anal. At. Spectrom.* **2006**, *21*, 847-855.

27. R. Djingova, H. Heidenreich, P. Kovacheva and B. Markert, *Anal. Chim. Acta*, **2003**, *489*, 245-251.

28. A. Ritter and V. R. Meyer, *Polym. Test.*, **2005**, *24*, 988–993.

29. K. Ivanov, Penka Zaprjanova, M. Petkova, V. Stefanova, V. Kmetov, D. Georgieva and V. Angelova, *Spectrochim. Acta B*, **2012**, *71-72*, 117-122.

30. W. Feller, "An Introduction to Probability Theory and Its Applications", 3rd edition, Wiley, New York, 1968.

31. J. Van Loco, M. Elskens, C. Croux and H. Beernaert, *Accred. Qual. Assur.*, **2002**, *7*, 281–285.

32. D. B. Hibbert, *Accred. Qual. Assur.*, **2005**, *10*, 300-301.

**Figure captions**

Figure 1. Response curve (broken line) may be approximated by the tangential line (dotted line) within the limits defined by eq. 2. A regression line (solid line) may be constructed and it may approximate well the response curve within the same range of concentrations when uncertainty of measurement is taken into account.

Figure 2. Response (y) of an apparatus depicted as a function of concentration (a) and( b), and the STDEV of y ($s_y$) according to the regression line of (b).

a) Tentative response curve (eq. 1) with measurements (square dots) calculated by MSExcel's random-number generator. b) Regression line obtained by the data of (a) within the range of concentrations given by eq. 2. c) STDEV ($s_y$) of (b) depicted as a function of concentration.

Figure 3. Experimental demonstration of law-of-large numbers. Determination by F-AAS of Co in certified references EUH-1 (a) and EPH-1 (b) where the average values ($\bar{x}$) are depicted as a function of number of repetitions (N). The corresponding REU's are depicted in c) and d), respectively. The results are shown consecutively in the order of measurements.

Figure 4. Illustration of non-correlation of result of unknown and slope within certain limits of uncertainty defined by an ellipse (dots) with axes that correspond to the REU of the two parameters ($\alpha$ and $\bar{x}$). Results were significantly different from consensus value (broken-line arrow) for EUH-1 (a) but excellent correspondence was found for EPH-1 (b).

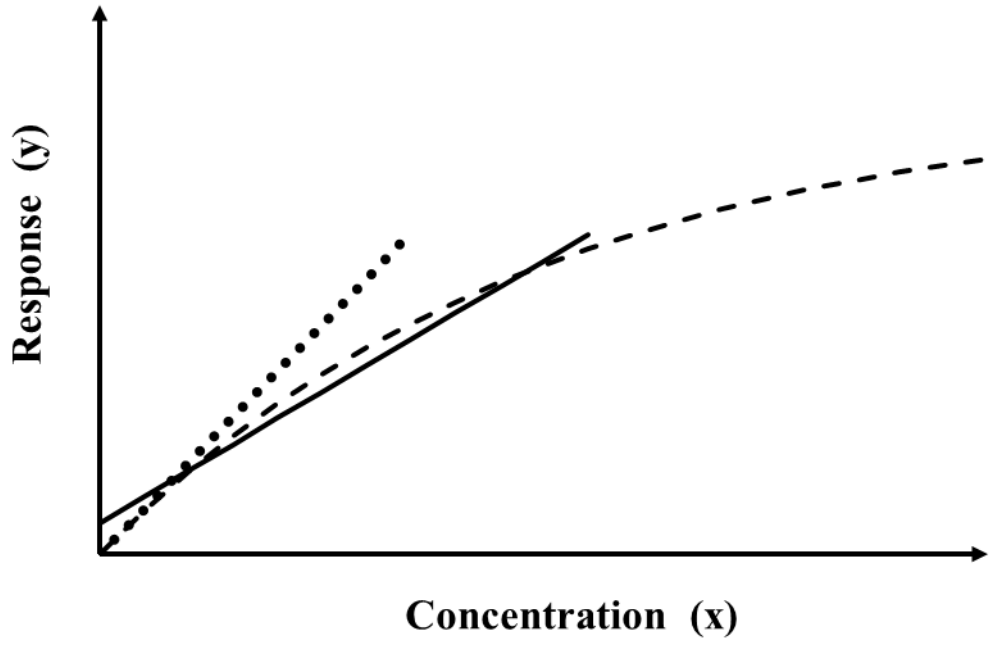


Fig. 1

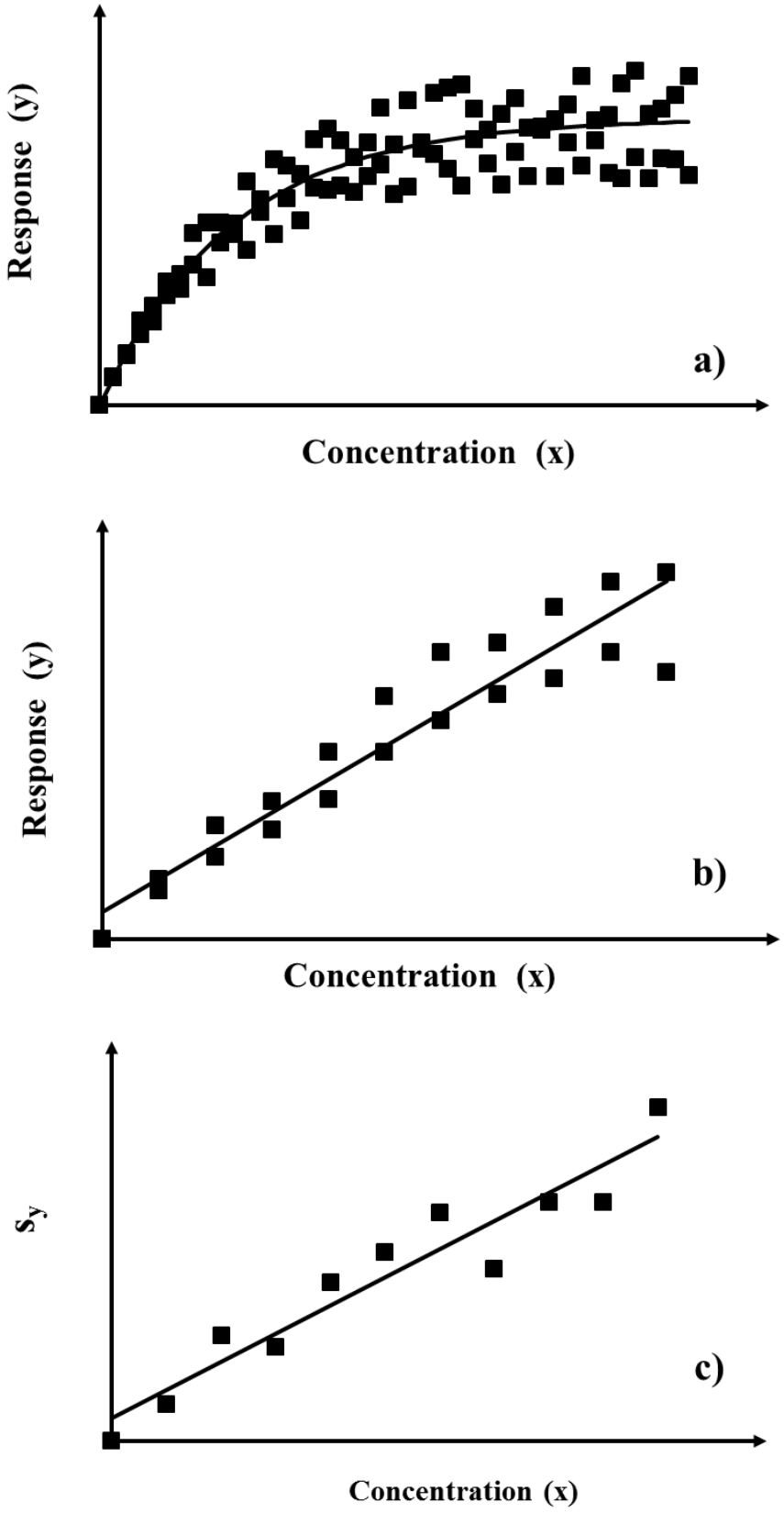
Response (y)
a)
Concentration (x)
Response (y)
b)
Concentration (x)
$s_y$
c)
Concentration (x)

Fig. 2

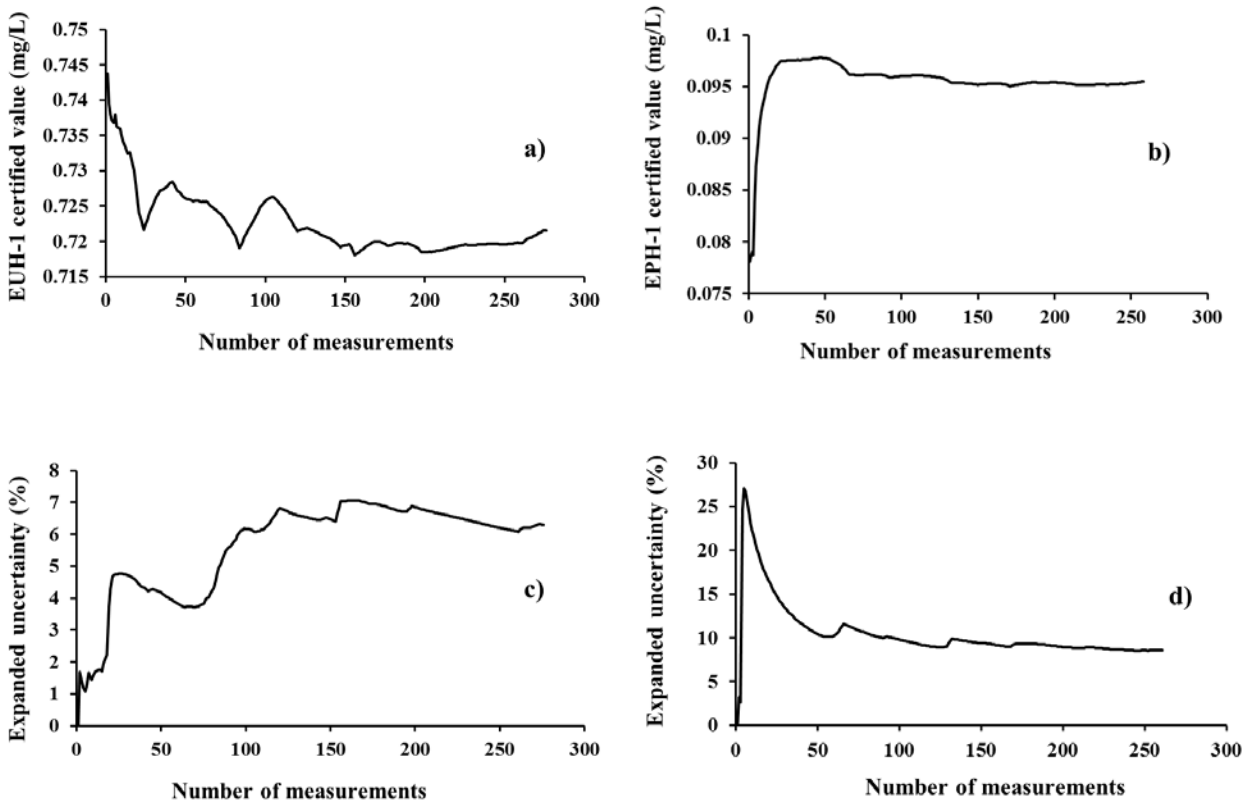
EUH-1 certified value (mg/L)
0.75
0.745
0.74
0.735
0.73
0.725
0.72
0.715
a)
0
50
100
150
200
250
300
Number of measurements
EPH-1 certified value (mg/L)
0.1
0.095
0.09
0.085
0.08
0.075
b)
Number of measurements
Expanded uncertainty (%)
8
7
6
5
4
3
2
1
0
c)
Number of measurements
Expanded uncertainty (%)
30
25
20
15
10
5
0
d)
Number of measurements

Fig. 3

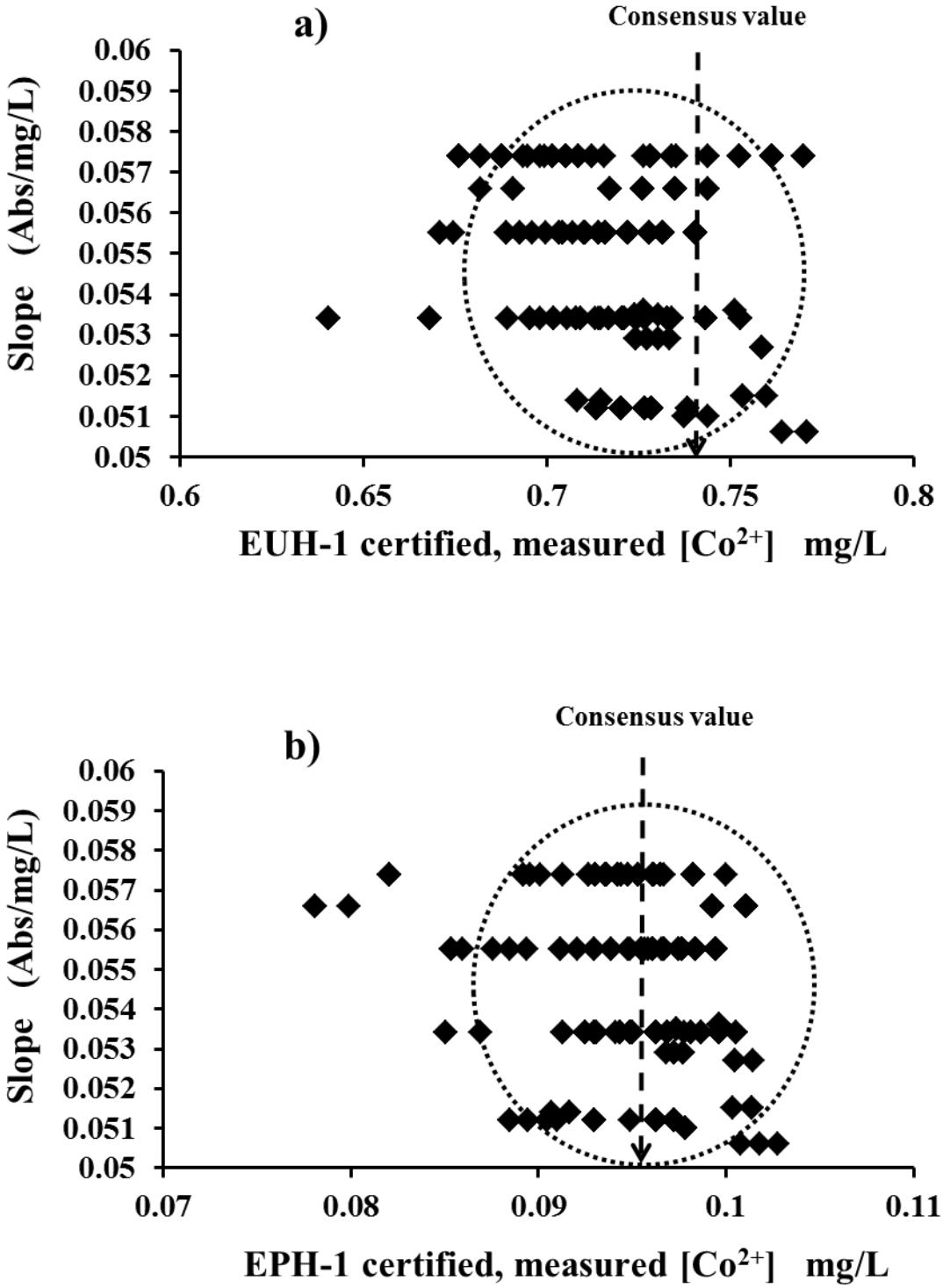
a)
Consensus value
Slope (Abs/mg/L)
0.06
0.059
0.058
0.057
0.056
0.055
0.054
0.053
0.052
0.051
0.05
0.6
0.65
0.7
0.75
0.8
EUH-1 certified, measured [Co2+] mg/L
b)
Consensus value
Slope (Abs/mg/L)
0.06
0.059
0.058
0.057
0.056
0.055
0.054
0.053
0.052
0.051
0.05
0.07
0.08
0.09
0.1
0.11
EPH-1 certified, measured [Co2+] mg/L

Fig. 4

Table 1

Concentrations of $Fe^{3+}$ in 0.1 M $HNO_3$ as determined by GF-AAS. Three possible calibration ranges obtained by iteration eliminating data of high concentrations within the non-linear-response range (eq. 1).[14] Intercept and blank value corresponded to each other at the third iteration. Abbreviations: Lower-limit of analysis (LLA), upper-limit of analysis (ULA), start-of-best range (SBR) and best-expanded uncertainty (BEU). The BEU was approximately constant and it did not improve by narrowing the calibration range.

| **Iteration #** | **Number of data for pooled-regression line** | **LLA (mg/L)** | **ULA (mg/L)** | **SBR (mg/L)** | **BREU (%)** | $\beta \pm 2 \cdot s_\beta$ **(Abs)** | $\bar{y}_0 \pm 2 \cdot s_{\bar{y}_0}$ **(Abs) (N = 10)** |
|---|---|---|---|---|---|---|---|
| 1 | 126 | 22 | 1300 | 131 | 12 | 0.096 ± 0.018 | 0.015 ± 0.020 |
| 2 | 116 | 14 | 820 | 60 | 17 | 0.068 ± 0.014 | 0.015 ± 0.020 |
| 3 | 77 | 7 | 300 | 30 | 16 | 0.0350 ± 0.0092 | 0.015 ± 0.020 |

Table 2

Figures of merits derived on the basis of pooled calibrations using 52 data points of 5 independent series of measurements. Relative-expanded uncertainties (REU's) are compared to those of the Horwitz formula.(ref.) Abbreviations: Lower-limit of analysis (LLA), start-of-best range (SBR) and best-relative-expanded uncertainty (BREU).

| **Compound** | **LLA (μg/L)** | **SBR (μg/L)** | **BREU (%)** | **REU(repetition) of unknown (N = 4)** | **REU(calibration) of unknown** | **REU of Horwitz formula** |
|---|---|---|---|---|---|---|
| 3-heptanol | 33 | 73 | 32 | 68 | 30 | 41 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Dimethyl sulphoxide | 31 | 76 | 28 | 88 | 34 | 37 |
| Butanoic acid | 31 | 48 | 46 | 112 | 44 | 35 |

Table 3

Determination by ICP-MS (pulse mode) of concentration of elements in certified reference (NIST Bone Meal 1486). Measurements were performed with $^{103}$Rh internal standard (IS) and the calculation of concentrations were performed both without IS and with IS. See text for abbreviations.

| m/z (amu/q) | Element | $\bar{x}$ ± EU (w/w%) N = 50 | Certified (w/w%) | LOQ** (µg/L) | LLA (µg/L) | ULA (µg/L) | SBR (µg/L) | BREU (%) | REU of CRM (%) | LLA IS (µg/L) | ULA IS (µg/L) | SBR IS (µg/L) | BREU IS (%) | REU IS of CRM (%) | $\frac{LLA}{LOQ}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | Na | 0.55 ± 0.10 | 0.50* | 0.03 | 14 | 500 | 26 | 40 | 20 | 17 | 500 | 26 | 48 | 72 | 510 |
| 24 | Mg | 0.440 ± 0.064 | 0.466 ± 0.017 | 0.03 | 27 | 800 | 27 | 44 | 14 | 20 | 800 | 27 | 52 | 20 | 600 |
| 25 | Mg | 0.440 ± 0.074 | 0.466 ± 0.017 | 0.03 | 71 | 5000 | 100 | 49 | 16 | 84 | 5000 | 100 | 58 | 18 | 2500 |
| 26 | Mg | 0.47 ± 0.15 | 0.466 ± 0.017 | 0.03 | 74 | 5000 | 100 | 51 | 30 | 88 | 5000 | 100 | 61 | 14 | 2600 |
| 31 | P | 12.8 ± 2.6 | 12.30 ± 0.19 | 3 | 38 | 5000 | 100 | 26 | 20 | 50 | 5000 | 100 | 59 | 14 | 17 |
| 42 | Ca*** | 27.7 ± 4.1 | 26.58 ± 0.24 | 0.3 | 110 | 5000 | 200 | 39 | 15 | 136 | 5000 | 264 | 36 | 16 | 420 |
| 43 | Ca*** | 19.0 ± 4.0 | 26.58 ± 0.24 | 0.3 | 67 | 5000 | 100 | 45 | 21 | 69 | 5000 | 110 | 46 | 20 | 210 |
| 88 | Sr | 276 ± 37 µg/g | 264 ± 7 µg/g | 0.003 | 3.7 | 200 | 11 | 24 | 14 | 2.4 | 200 | 16 | 10 | 14 | 720 |

*Indicated value **Manufacturer's specifications ***Determined at concentrations much above ULA