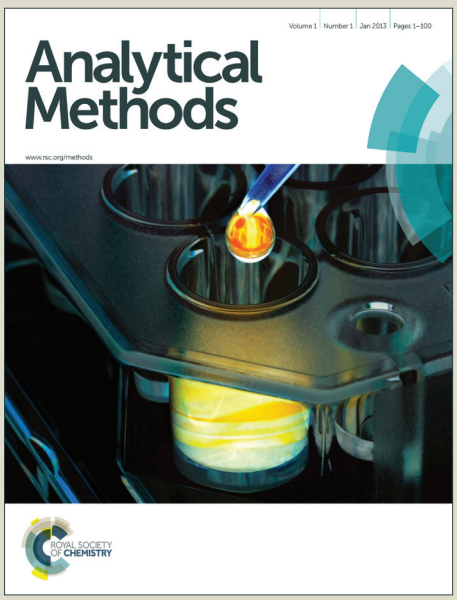
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Proficiency testing for determination of lead and arsenic in cosmetics: Comparison of analytical procedures and evaluation of laboratory performances

Zhixiong Zhong [a*], Gongke Li [b], Jianbo Luo [a], Wensheng Chen [a], Liping Liu [a], Ping He [c], Zhibin Luo [a]

[a] *Center for Disease Control and Prevention of Guangdong Province, Guangzhou 510300, China;*

[b] *School of Chemistry and Chemical Engineering, Sun Yat-sen University, Guangzhou 510275, China*

[c] *China National Accreditation Service for Conformity Assessment, Beijing 100062, China*

*Corresponding author: Z. X. Zhong

Tel.: +86 20 31051925

Fax: +86 20 31051925

E. mail address: zzxiong608@126.com

Abstract

A proficiency testing scheme (CNAS T0419) concerning lead and arsenic determination in foundation cream cosmetics involved 217 laboratories from China as participants, who used their regular analytical methods. The metrological comparability of results from various approaches was tested using simple linear regression analysis and analysis of variance that were valid techniques. There were no significant differences in values after elimination of outliers by the Grubbs test. The normality of the distribution of results submitted by participants was examined with the one-sample Kolmogorov-Smimov test. The distribution characteristics of result points for the paired samples were investigated using the Youden plot, which could greatly simplify data analysis process. The assigned values of test materials for proficiency assessment were derived directly from the reported results of the participating laboratories. The measurement capabilities of laboratories for analyzing heavy metals, especially lead and arsenic in complex matrices were objectively assessed using *z*-scores, which could readily be used to compare each participant's value relative to the others and would enable the outlier laboratories to find out the dominant error sources such as systematic and random variations. The percentages of acceptable results of between- and within-laboratories were 86.0%, 84.2% for lead, and 84.6%, 83.6% for arsenic, respectively.

Keywords: Proficiency testing, Youden plot, *z*-scores, lead, arsenic, cosmetics

1. Introduction

The beauty consciousness of people promotes the development of cosmetics, which its ingredients are further complex to achieve specific functions. These ingredients including natural and synthetic chemical substances are directly applied to human skin, oral cavity and mucosa. The risk of allergic dermatitis maybe increases owing to the long-term exposure to these ingredients in normal use. Heavy metals like lead (Pb) and arsenic (As) may occur in cosmetics as a result of inadvertent migration from the metallic devices used during the manufacturing process and the use of poor quality raw ingredients. The toxic heavy metals can cause serious health hazards such as acute and chronic poisoning, diverse diseases and even lead to cancer after accumulating in human body [1, 2].

The occurrence of heavy metal contamination in cosmetics has attracted worldwide attention in recent years. The heavy metal in cosmetic raw materials and finish products have been strictly monitored and controlled by the local authorities. Even if trace metals in cosmetics can also be detected using modern specific instrument. As early as 2007, the Campaign for Safe Cosmetics has reported that a large number of popular brands of lipsticks and lip glosses contained lead [3]. In 2009 and 2011, the United States Food and Drug Administration (US FDA) had found lead in many commercial lipsticks at levels much higher than those values detected by the Campaign for Safe Cosmetics [4]. Subsequently, a survey concerning the 15 Member States of European Union indicated that 31% of the tested lipsticks and 4% of the lip glosses contained lead at a level higher than 1 mg kg^{-1} [5]. The China Food and Drug Administration (CFDA) had notified that the contents of lead, arsenic

and mercury in certain world-famous cosmetic brands exceeded the maximum authorized concentration in the finished products [6]. Several eye shadow cosmetics from various countries have been discovered to contain dangerous heavy metals at high levels [7]. The eight kinds of banned toxic metals in cosmetics have been specifically reviewed [8]. These cosmetics safety issues have already caused serious impacts on international trade and manufacture of cosmetic products.

Several toxic elements including lead, arsenic, cadmium, antimony, chromium, cobalt, mercury, and nickel are listed in the Annex II of European Cosmetic Regulation 1233/2009. These metals must not form parts of the composition of cosmetic products because of their potential toxicity [9]. Many countries have formulated strict limits for Pb and As as impurities in cosmetics. The US FDA legislation requires that the limit for Pb as impurities in color additives is 20 mg kg^{-1} [10]. While the CFDA limits are 40 mg kg^{-1} for Pb and 10 mg kg^{-1} for As in diverse cosmetic products [11]. Various approaches are available for the analysis of Pb and As in cosmetics [12]. The extracted amounts of Pb and As depend completely upon sample preparation procedures and experimental conditions used due to the complex cosmetic matrices. Sample treatment is crucial to analyzing heavy metals in cosmetics. The majority of procedures can achieve complete digestion of the cosmetic matrices using the mixtures of concentrated acids (e.g. nitric acid, sulfuric acid and perchloric acid) with hydrogen peroxide (oxidant) in an open system at high temperature. To accelerate dissolution and digestion of complex matrix small amounts of hydrofluoric acid (about 1mL) is usually added to polytetrafluoroethylene tank. The use of hydrofluoric acid should pay attention to safety due to its strong corrosivity. Microwave digester and high pressure vessel are commonly utilized

to improve digestion efficiency and specificity [13-16]. The quantitative techniques for the determination of Pb mainly include the inductively coupled plasma optical emission spectrometry (ICP-OES) [17], the inductively coupled plasma atomic emission spectrometry (ICP-AES) [18], the inductively coupled plasma mass spectrometry (ICP-MS) [5], the flame atomic absorption spectrometry (F-AAS) [19, 20], and graphite furnace atomic absorption spectrometry [21]. Several methods have been described for the determination of As such as the hydride generation atomic absorption spectrometry (HG-AAS) [22], ICP-OES [17] and ICP-MS [23]. Additionally, the instrumental neutron activation analysis (INAA) [24] can be directly applied to solid cosmetics to avoid the laborious sample preparation process.

In the last decade private and government testing laboratories in China have drastically increased for the analysis of toxic element and compounds in cosmetic raw materials and products. Although some of laboratories have been accredited, the comparability of analysis results from different laboratories is never assessed using practical techniques. Various analytical methods were utilized for determining Pb and As in cosmetics, the “fitness for purpose” of these procedures, and the accuracy and reliability of measurement results become key issues. The performance characteristic of laboratory may be demonstrated through interlaboratory comparison exercises or proficiency testing (PT) campaigns [25, 26]. The PT is known to act as a unique external control for the quality assurance and an essential tool for assessing the technical competence of participating laboratories.

In this study, a PT scheme (CNAS T0419) was designed in accordance with the general requirements of the ISO/IEC 17043 [27] and CNAS-RL02 [28] and organized in collaboration with the China National Accreditation Service (CNAS) for Conformity Assessment.

Numerous cosmetic testing laboratories from China subscribed for participation in the PT program. The participants were required to test the contents of Pb and As in foundation cream samples with complex matrices at two levels by using their analysis methods. The proficiency of participant and the comparability of results from different procedures were evaluated through statistical analysis. The distribution characteristics of the reported results were investigated with the Youden two-sample plots.

2. Experimental

2.1. Preparation of test materials

An ideal test material for a PT program would fulfill certain criterias, such as stable during material storage and transport to the participating laboratories, homogeneous across all the aliquots of materials produced, appropriate analyte concentration and matching matrix as close as possible to the real sample, available in sufficient volume and easy to transport. In practice, it is impossible to achieve all these goals, and some compromises are required in the preparation of test materials for PT. A native Pb and As in cosmetics can seldom provide a high enough concentration for a PT scheme, therefore, most of the test materials involve the spiking of the blank sample with standard solution.

A foundation cream was supported by a cosmetic manufacturer in Guangzhou, China. Its major ingredients include cyclopentasiloxane, ethylhexyl methoxycinnamate, dimethicone, glycerin, C30 alkyl ceteryl dimethicone crosspolymer, isopropyl isostearate, zinc oxide, dimethiconol, silica, PEG-20 dimethicone, propylparaben, propylene dlycol, trisodium-EDTA,

triethoxycaprylylsilane and fragrance. These materials were initially screened for the presence and concentration of Pb and As to accurately control the spiking level. The spiked concentration was controlled according to the CFDA limits for lead and arsenic as impurities in cosmetics. After being added with known amounts of Pb and As standard substances, the samples were fully blended for 10 h using a high-speed homogenizer at a moderate speed varying from 500 to 800 rpm under the controlled temperature of 35 ^{O}C, and then left overnight to achieve complete interaction between analytes and matrices. Aliquots of about 50 g of the well-mixed foundation cream samples were packed into pre-cleaned plastic bottles with screw cape and sealed using polypropylene bags. More than 250 packaged units at each level were prepared and stored at room temperature prior to dispatching to participants.

2.2. Homogeneity and stability of test materials

Homogeneity test is a crucial step that validates whether the material will be suitable for the use of proficiency testing. All test materials underwent homogeneity testing for analytes prior to dispatch. The number of units examined is necessarily small to keep the cost of the test within bounds [29]. The contents of Pb and As in ten randomly selected subsamples were analyzed in duplicate using ICP-MS methods. To determine the within- and between-bottle variations two aliquots of 0.5 g from each of the ten random subsamples were placed into twenty 50-mL screw-capped polyethylene tubes, and 8 mL of 65% (w/w) nitric acid and 3 mL of 30% (w/v) hydrogen peroxide were added. These samples were digested for 24 h in high pressure vessels at temperature of 180 ^{O}C. The digested sample solutions were heated to remove NO_x in water bath at 100 ^{O}C, and then diluted with deionized water. Sample homogeneity was assessed according to the ISO/IEC 17043 and CNAS-RL02. The results of

one-way analysis of variance (ANOVA) confirmed that these samples were sufficiently homogeneous (P>0.05).

The stability study of the test materials prepared was conducted using the similar protocol at the beginning, during storage, and after the receipt of test results from the participants. The average concentrations of analytes in three random samples for duplicate analysis at each storage interval were compared with the mean values from the homogeneity test using the *t*-test. There were no statistically significant differences of these results, and the differences between test values were no more than 30% of σ_R (σ_R was estimated using the Horwitz function of $0.02\,c^{0.8495}$, where c was the consensus value or assigned values), indicating that these samples were stable enough in the life-time and suitable for the PT samples. The program would provide Pb and As residue analysis in samples at four concentration levels. These four samples with similar matrices, including sample code 1 A, 1 B, 2 A and 2 B, were divided into two groups (the former two and the latter two samples). One group of paired-samples was dispatched to NO.1-100 participants, and the other one was distributed to NO.101-217 laboratories after numbering at random in August 2009.

2.3. Performance assessment

All data received from the participants were evaluated using a standard procedure that allowed the direct comparison among values. The assigned value and the participants results were assessed with robust statistics, which is generally used to calculate a consensus value without the need for eliminating outliers from the raw data set, may weaken the effects of the method on the results, and provides a more reliable estimate of the measurement relating to the procedure [30]. The robust statistics can routinely cope with near-normal distribution data

and achieve a perfect estimation of the consensus values.

Participants were required to perform two independent measurements for each of paired samples and to report the mean using their preferred analytical methods. The statistical analysis of the results submitted by participants was carried out employing the Microsoft Office Excel 2010 and Statistical Package for the Social Science (SPSS) version 16.0. The statistic parameters involved number of results, median, interquartile range (*IQR*), normalized interquartile range (*NIQR*), *z*-score between-laboratories (ZB) and *z*-score within-laboratories (ZW), the values of which were calculated by using the following formulas: $IQR=Q_3-Q_1$, $NIQR=0.7413\ IQR$, $ZB=\left|S - median\ (S)\right| / NIQR\ (S)$ and $ZW=\left|D - median(D)\right| / NIQR(D)$, where S (standardized sum)$=(A+B)/\sqrt{2}$, D (standardized difference) $= (A-B)/\sqrt{2}$. The performance of participants was evaluated according to the generally accepted limits [27, 28]: $|ZB|\leq 2$ and $|ZW|\leq 2$, satisfactory result; $2<|ZB|<3$ or $2<|ZW|<3$, questionable result; $|ZB|\geq 3$ or $|ZW|\geq 3$, unsatisfactory result.

3. Results and discussion

3.1 Assigned values

Assigned value affects the validity of performance assessment in PT schemes. Various procedures can achieve the assigned values. A common procedure is to categorize participant approaches into "peer groups" that represent similar technology and calculate the mean or median of the peer group as the assigned value after the elimination of outliers [31]. However, as the number of participants decreases or the dispersion among the results increases, the

systemic error of the assigned value increases. It seems that an assigned value obtained from a reference measurement procedure or mean (or median) of all-methods is satisfactory [32, 33]. Such approaches are a lack of scientific rigor and sometimes the values of the former are likely larger than those of the latter [34]. The assigned value was usually given by the original Horwitz function in the United Kingdom Food Analysis Performance Assessment Scheme [35]. While a consensus value derived directly from the results of participants was used as the assigned value in European Commission proficiency testing schemes (EUCPTs) for pesticide residues analysis in fruits and vegetables. The reported results had not been adjusted to a normal distribution, the central value was considered to be the robust median (assigned value) [36]. This assessment procedure has widely adopted by PT providers due to its statistical validity. In the PT scheme, the test materials were analyzed using diverse approaches in different laboratories including organizer, expert laboratory and participants. The overall data, displayed in Table 1, suggest that the results were excellent agreement with each other in a peer group. The homogeneity mean / median ratio values were close to 1.0 with the mean ratio of 0.9995, likewise the expert laboratory mean / median ratio values ranged from 0.976 to 1.013 with the average value of 0.992, and there seems to be no concentration-related change in the ratio values. Hence the robust median value can be used as the assigned reference value for proficiency test.

3.2. Assessment of metrological comparability of results

A total of 217 laboratories from 31 provinces, municipalities and autonomous regions in China, registered for the PT program and 207 of them returned their results punctually to the organizers. Results for both Pb and As in test materials were reported by 202 laboratories,

three for only Pb and two for As. Although there were a number of results that were outside the whisker limit (1.5 times IQR) it was desirable to keep all but the most extreme outliers. Hence, the well known Grubbs test was employed to justify the abnormally high deviations from the mean of laboratories [37]. As can be seen in Table 2, microwave-assisted digestion was the most favourable choice in sample preparation for analysis of Pb (49.9%) and As (46.8%) in test materials. Wet digestion was adopted in the second procedure and its percent was 42.3% for Pb and 46.6 % for As. This procedure can be utilized to completely digest all kinds of matrices including lipin, wax and mineral substances. In addition to these two procedures, dry ashing and acid extraction techniques were also employed by a few of participants.

The simple linear regression analysis can examine the comparability of results obtained from various approaches [38]. The relationship between the mean values (y) for each of sample preparation procedures and the assigned value (x) of test material was linear. Their slopes of the linear regression were in the range of 0.985-1.013, which were close to 1.0. This implies that there were no obviously analytic biases among procedures (Fig.1 a). The results gained from diverse sample preparation techniques were fully comparable, and there are no statistically significant differences among them (P-value > 0.05 and $F<F$ critical) after elimination of outliers by the Grubbs test. In fact, there were a few of outlier values in the use of dry ashing procedure due to the existence of interference or loss of analyte.

After being digested, two elements in solution were determined by using FASS (90.2%), ICP-OES (5.1%) and ICP-MS (4.7%) for Pb, and HG-AAS (88.1%), ICP-OES (4.9%), ICP-MS (1.5%) and spectrophotometry (5.5%) for As. An external standard calibration

method was mainly employed for the quantification of analytes. The results of participants were classified to calculate the mean values and compare analytical variances according to the procedures used. Fig.1 b show that regression line slopes of the mean values (y) for each of measurement approaches versus the assigned value (x) ranged from 0.961 to 1.022 with the average value of 0.995. There was metrological comparability among these approaches. A previous report [39] revealed that the analytical results of Pb and Cd in workplace were more superior when using the ICP-OES/ICP-AES than the FASS, but no significant difference (p>0.05) was observed in the PT scheme (Table 3). The outcomes were in good agreement with those of the Asia Pacific Laboratory Accreditation Cooperation (APLACT) T065 program for the analysis of Pb and Cd in herbal material [39]. Hence, the large deviation results of participants should not be attributed to the different sample treatment and detection techniques.

3.3. Evaluation of laboratory proficiency

3.3.1. Youden plots

All data were combined according to peer group for the statistical evaluation after collection. The distributions of result points for the paired samples were displayed in the Youden plots (Fig. 2 a-d). Confidence ellipse region (α=0.05) for the two-sample plot was expressed in terms of the Hotelling's T^2 distribution [40]. The majority of data points were concentrated at the center of the ellipse. This demonstrated that performances of participants for both samples were satisfactory. A few of data points were located outside of the ellipse, where it didn't mean that all these corresponding participants had unsatisfactory results. Because the critical value of $|z|=3$ is associated with coverage probability of 99%, in most

cases (95%), more than 5% of points should fall outside of the ellipse due to the existence of outliers. Therefore, certain questionable values also scattered around the ellipse. The data points (square) with unsatisfactory results ($|z| \geq 3$) were labeled as laboratory codes to distinguish them from the questionable values outside of the ellipse. Among the paired samples the highly-dispersive values were drived from the Pb in 2 A and 2 B samples (Fig. 2 c) due to the existence of the relatively large systematic and random errors in some laboratories. The majority of data points of As in 2 A and 2 B samples (Fig. 2 d) were centered in the ellipse with high accuracy. Several data points including laboratories codes 27, 76, 180 and 194 were in the lower left quadrant where their values for the paired samples were very low. In contrast, certain data points (codes 5, 67, 175 and 186) were in the upper right quadrant with the extremely high values. These suggested that large systematic errors (between-laboratories variation) occurred in these laboratories. A few of laboratories (codes 62, 209 and 188) were located in the upper left quadrant where they had high values for both samples, while one laboratory (codes 71) were fallen in the lower right quadrant with much smaller results than the assigned values. There were the substantially random errors (within-laboratories variation) in these four laboratories. There was probably the large bias in spectrophotometry due to the existent of matrix effect (codes 62 and 175). The complex test materials could not be completely digested using microwave-assisted digestion in usual conditions, and the organic residues would affect accurate measurement of Pb and As in digestive solution (e.g. codes 5, 27, 71, 180, 188 and 209). The percentages of unsatisfactory results were 12.1% for Pb, 15.2% for As in 1 A and 1 B, and 14.7% for Pb, 11.3% for As in 2 A and 2 B samples. It should be highlighted that the dispersion of results didn't depend on

the analyte and its concentration, and the methods used.

3.3.2. Comparison of ZB- and ZW-score values

Quantitative results are usually valuated by means of the robust *z*-score. Raw data were not all normally distributed, but the histograms illustrate that their modified data sets were closely approximated a normal distribution after exclusion of outliers (Fig. S1. a-d of the supporting information). The results of nonparametric test of the reported values for each analyte in samples were listed in Table S1 (the supporting information), indicating that these data followed a normal distribution. The mean value for each of samples was in good agreement with its robust median. The robust median and *NIQ*R (0.7413*IQR*) should be used to convert the participants' results into *z*-scores for assessing their performances in the PT program.

Several statistics parameters including number of results, range of results, medians, robust coefficient of variation (CV), number and percentage of satisfactory and unsatisfactory results are generally recommended to explain PT data by National Association of Testing Authorities (NATA). The results of robust statistics were summarized in Table 4. The total satisfactory rates were in the range of 81.9%-88.8% for Pb and 79.8%-87.0% for As. The percentages of acceptable results of between- and within-laboratories were 86.0%, 84.2% for Pb, and 84.6%, 83.6% for As, respectively. These findings are consistent across the study of international proficiency testing program (APLAC T 065) on cadmium and lead in herbal sample [40].

The overall performance of participant was comprehensively assessed using two-type *z*-scores (Fig. 3 a-d). The majority of laboratories performed well with $|ZB| \leq 2$ and $|ZW| \leq 2$.

But four laboratories (codes 67, 106, 175 and 186) overestimated their results with the extremely positive ZB and ZW values. Contrary to this case, another four laboratories (codes 27, 76, 180 and 194) underestimated their results and had the extremely negative ZB and ZW values. All of these results were significantly deviated from the assigned values due to the existence of between- and within-laboratories measurement errors. The most probable reason was that the sample pretreatment step led to high variability in the results obtained. Because these complicated test materials contain a lot of lipid and inorganic salts that can't be effectively digested using oxidizing acid at a not high enough temperature, arsenic in material can't translate completely into free ions, resulting in an underestimation of values using HG-AAS measurement (e.g. codes 27 and 194). The residual NO_x in test solution may cause an overestimation of results. Additionally the existent of matrix effect can seriously affect test results (e.g. codes 67 and 86). It was beyond doubt that these laboratories given the extreme values should be performed an in-depth investigation for the entire operation procedures and taken some effective measures to correct the deviations.

4. Conclusion

A PT scheme was implemented using paired samples for Pb and As residue analysis in cosmetic materials. The simple linear regression analysis and ANOVA were used to examine the comparability of results obtained from diverse methods. The results demonstrated that there was no significant method-dependence for the analysis conducted in the PT scheme. The performances of participants were assessed with two kinds of techniques including the

Youden plot and *z*-score. The random distribution characteristics of the data sets and the outlier values from the participants might readily be observed in the Youden plots. This semi-quantitative evaluation technique had no need to use the complex statistical technique for the conclusion, and it was extremely useful to simplify data analysis process. The mean values were in good accord with the assigned values after elimination of outliers by the Grubbs test, revealing that the majority of participants had excellent measurement capability for Pb and As in complicated cosmetic materials. The ordered ZB- and ZW-score charts were easily used to compare each value of participant relative to the others, which would enable the outlier laboratories to find out the main error sources as a result of a specific procedure or a given laboratory and help to comply with quality assurance and quality control requirements. The absence of general good analytical practice is a more likely factor in poor performance.

Acknowledgement

The authors would like to express their gratitude to the China National Accreditation Service for Conformity Assessment for providing resources to this scientific project. We appreciate the assistance from Hongjie Liang, Jing Wang, Zhifeng Li and Yijuan Shao.

References

1 L. Jarup, *Br. Med. Bull*., 2003, **68**, 167-182.

2 K. Koller and T. Brown, A. Spurgeon, L. Levy, *Environ. Health Perspect.,* 2004, **112**, 987-994.

3 FDA, *Lipstick and lead: Questions and answers (website)*, Silver Spring, MD: U.S. Food and Drug Administration (updated 7 Dec 2012), available: http://goo. gl/qT 5 mm (accessed 7 May 2013).

4 The Campaign for Safe Cosmetics, *A poison kiss: The problem of lead in lipstick, Safe Cosmetics Action Network (Oct 2007),* available: http://goo.gl/71qW9 (accessed 7 May 2013).

5 P. Piccinini, M. Piecha and S. F. Torrent, *J. Pharm. Biomed. Anal.*, 2013, **76**, 225-233.

6 China Food and Drug Administration, available: http://www. sda. gov. cn/ WS01/ CL0051/ 98134. html, 2013.

7 M. G. Volpe, M. Nazzaro, R. Coppola, F. Rapuano and R. P. Aquino, *Microchem. J.*, 2012, **101**, 65-69.

8 B. Bocca, A. Pino, A. Alimonti and G. Forte, *Regul. Toxicol. Pharm.,* 2014, **68**, 447-467.

9 European Commission, *Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products*, *O. J.* L 342, 22. 12. 2009, pp. 59-209.

10 I. Al-Saleh, S. Al-Enazi and N. Shinwari, *Regul. Toxicol. Pharm.,* 2009, **54**, 105-113.

11 Ministry of Health of the People's Republic of China. *Hygienic standard for cosmetics*, Beijing, China, 2007, pp. 166-179.

12 ISO/TR 17276, *Analytical approach for screening and quantification methods for heavy metals in cosmetics*, International Organization for Standardization, Genève, Switzerland, 2014.

13 A. A. Adepoju-Bello, O. O. Oguntibeju, R. A. Adebisi, N. Okpala and H. A. B. Coker, *Afr.*

J. Biotechnol., 2012, **11**, 16360-16364.

14 N. Hepp, *J. Cosmet. Sci.,* 2012, **63**, 159-176.

15 N. R. Cha, J. K. Lee, Y. R. Lee, H. J. Jeong, H. K. Kim and S. Y. Lee, *Anal. Lett.*, 2010, **43**, 259-268.

16 B. Bocca, G. Forte, A. Pino and A. Alimonti, *Anal. Methods,* 2013, **5**, 402-408.

17 S. Liu, K. Hammond and A. Rojas-Cheatham, *Environ. Health Perspect.*, 2013, **121**, 705-710.

18 A. A. Alqadami, M. A. Abdalla, Z. A. Alothman and K. Omer, *Int. J. Environ. Res. Public Health,* 2013, **10**, 361-374.

19 E. O. Amartey, A. B. Asumadu-Sakyi, C. A. Adjei, F. K. Quashie, G. O. Duodu and N. O. Bentil, *Br. J. Pharmacol. Toxicol.*, 2011, **2**, 192-198.

20 G. D. Tarigh and F. Shemirani, *Talanta*, 2013, **115**, 744-750.

21 R. F. de Oliveira, C. C. Windmőller, W. B. Neto, C. C. de Souza, M. A. Beinnerd and J. B. B. da Silv, *Anal. Methods,* 2013, **5**, 5746-5752.

22 A. N. Anthemidis, E. I. Daftsis and N. P. Kalogiouri, *Anal. Methods,* 2014, **6**, 2745-2750.

23 J. C. Raposo, P. Navarro, J. I. G. Felipe, J. Etxeandia, J. A. Carrero and J. M. Madariaga, *Microchem. J.*, 2014, **114**, 99-105.

24 L. Sneyers, L. Verheyen, P. Vermaercke and M. Bruggemann, *J. Radioanal. Nucl. Chem.*, 2009, **281**, 259-263.

25 E. Pereira, S. M. Rodrigues, M. Otero, M. Válega, C. B. Lopes, P. Pato, J. P. Coelho, A. I. Lillebø, M. A. Pardal, R. Rocha and A. C. Duarte, *TrAC Trends Anal. Chem.,* 2008, **27**, 595-570.

26 ISO/IEC 17025, *General requirements for the competence of testing and calibration laboratories*. International Organization for Standardization, Genève, Switzerland, 2005.

27 ISO/IEC 17043, *Conformity assessment-general requirements for proficiency testing*. International Organization for Standardization, Genève, Switzerland, 2010.

28 CNAS-RL02, *Rules for proficiency testing*. China National Accreditation Service for Conformity Assessment, Beijing, China, 2007.

29 M. Thompson and T. Fearn, *Anal. Methods*, 2011, **3**, 2529-2533.

30 W. P. Collno, I. H. M. Van Stokkum, I. Van Steenwijk and D. F. Wells, *Anal. Chim. Acta*, 2005, **533**, 31-39.

31 W. G. Miller, G. R. D. Jones, G. L. Horowitz and C. Weykamp, *Clin. Chem*., 2011, **57**, 1670-1680.

32 M. Heuillet, B. Lalere, M. Peignaux, J. De Graeve, S. Vaslin-Reimann, J. P. P. De Barros, P. Gambert, L. Duvillard and V. Delatour, *Clin. Biochem.*, 2013, **46**, 359-364.

33 J. Á. Rodríguez-Castrillón, M. Moldovan and J. I. G. Alonso, *J. Anal. At. Spectrom.*, 2009, **24**, 815-824.

34 T. Yarita, T. Otake, Y. Aoyagi, T. Kuroiwa, M. Numata and A. Takatsu, *Talanta*, 2015, **132**, 269-277.

35 M. Thompson, M. Sykes, M. Knaggs and S. Hunter, *Anal. Methods*, 2013, **5**, 4927-4928.

36 P. Medina-Pastor, C. Rodríguez-Torreblanca, A. Andersson and A. R. Fernández-Alba, *TrAC Trends Anal. Chem.,* 2010, **29**, 70-83.

37 ISO 5735-2, *Accuracy (trueness and precision) of measurement methods and results. Part 2. Basic method for the determination of repeatability and reproducibility of a standard*

measurement method, International Organisation for Standardisation, Genève, Switzerland, 1994.

38 B. W. Adam, E. M. Hall, N. K. Meredith, T. H. Lim, C. A. Haynes, V. R. De Jesus and W. H. Hannon, *Clin. Biochem.*, 2012, **45**, 1658-1663.

39 P. Y. T. Hon, P. K. Chan, S. T. C. Cheung and Y. C. Wong, *Microchem. J.*, 2011, **98**, 44-50.

40 D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi and J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: part A*, Elsevier, Amsterdam, NL, 1997, pp. 452-454.

Figure Captions

Fig. 1. Linear regression analyses of results obtained from different procedures. (a) Pb: microwave-assisted digestion; y=1.012x-0.59, r=0.9998; wet digestion, y=1.012x-0.47, r=0.9999; acid extraction, y=1.036x-0.77, r=0.9998; As: microwave-assisted digestion, y=0.985x+0.10, r=0.9996; wet digestion, y=0.995x+0.004, r=0.9998; dry ashing, y=0.991x-0.13, r=0.9982; (b) Pb: ICP-OES, y=0.979x-0.07, r=0.9994; ICP-MS, y=1.010x-0.64, r=0.9986; FASS, y=1.022x-0.71, r=0.9999; As: ICP-OES, y=0.961x+0.24, r=0.9978; HG-AAS, y=0.992x+0.04, r=0.9997; spectrophotometry, y=1.003x-0.07, r=0.9998.

Fig. 2. Youden two-sample plots of concentrations for (a) Pb in samples 1 A and 1 B, (b) As in samples 1 A and 1 B, (c) Pb in samples 2 A and 2 B, (d) As in samples 2 A and 2 B. The *X*-axis and *Y*-axis indicated the average concentrations (mg kg^{-1}) of Pb or As in A and B samples. The critical points ($|z|$=3) were plotted as ellipse at a confidence level of 99%.

Fig.3. Participants' ZB- and ZW-score values for (a) Pb in samples 1 A and 1 B, (b) As in samples 1 A and 1 B, (c) Pb in samples 2 A and 2 B, (d) As in samples 2 A and 2 B. Dotted and solid lines correspond to $|z|$=2 and $|z|$=3, respectively.

Table 1

Results of statistics analysis of Pb and As concentration (mg kg^{-1}) in test materials

Analyte	Pb				As			
Sample	1 A	1 B	2 A	2 B	1 A	1 B	2 A	2 B
Mean value ±*SD*(*n*)[a]	15.2±0.5(97)	35.4±0.95(94)	17.9±1.2(114)	30.2±1.2(112)	3.23±0.27(97)	8.21±0.47(91)	3.59±0.19(110)	7.17±0.39(110)
Median ±*NIQR* (*N*)[b]	15.2±0.4(99)	35.4±0.9(99)	18.0±1.0(116)	30.2±0.9(116)	3.24±0.20(99)	8.32±0.36(99)	3.60±0.12(115)	7.16±0.27(115)
Homogeneity mean ±*SD*[c]	15.5±0.4	35.7±0.8	17.8±0.5	30.1±0.6	3.27±0.12	8.20±0.15	3.65±0.18	7.08±0.22
Expert laboratory mean ±*SD*[d]	15.4±0.5	35.5±0.9	17.9±0.7	29.8±0.6	3.21±0.10	8.18±0.11	3.56±0.13	6.99±0.25
Homogeneity mean / median value	1.020	1.008	0.989	0.997	1.009	0.986	1.014	0.989

Expert laboratory mean/ median value	1.013	1.003	0.994	0.987	0.991	0.983	0.989	0.976

[a] Mean value, standard deviation (SD) and number (*n*) of the reported results after removal of outliers by the Grubbs test,

[b] Median value and normalized interquartile range of all the results (the total number) with robust statistics,

[c] Mean value and standard deviation of the test results (*n*=20),

[d] Mean value and standard deviation of ten replicates.

Table 2

Comparison of results obtained from diverse sample preparation procedures for the determination of Pb and As in test materials

Sample	Analyte	*F*	*F* critical	*P*-value	Microwave-assisted digestion		Wet digestion		Dry ashing		Acid extraction	
					n [a]	Mean±SD [b]	*n*	Mean±SD	*n*	Mean±SD	*n*	Mean±SD
1 A	Pb	0.9755	3.0912	0.3807	55	14.6±0.4	35	14.8±0.5	- [c]	-	9	14.8±0.6
	As	1.8699	3.0943	0.1599	45	3.32±0.31	46	3.21±0.22	5	3.24±0.20	-	-
1 B	Pb	1.1956	3.0912	0.3070	55	35.1±1.7	35	35.4±1.3	-	-	9	35.9±1.6
	As	1.3799	3.0933	0.2566	45	8.23±0.55	46	8.23±0.57	5	8.39±0.37	-	-
2 A	Pb	0.1406	3.9352	0.1406	50	17.8±1.6	53	17.9±1.1	-	-	9	18.1±0.8
	As	0.0502	3.0796	0.9510	54	3.59±0.23	51	3.59±0.24	7	3.56±0.12	-	-
2 B	Pb	0.0984	3.9343	0.7544	51	30.1±2.0	53	30.0±1.3	-	-	9	30.5±1.8
	As	1.9783	3.0796	0.1432	54	7.24±0.48	51	7.20±0.43	7	6.88±0.27	-	-

[a] *N*umber of the reported results after elimination of outliers by the Grubbs test,

[b] Units are mg kg^{-1} for mean value and SD,

[c] Not detection

Table 3

Comparison of results obtained from various measurement techniques for the analysis of Pb and As in test materials

Sample	Analyte	*F*	*F* critical	*P*-value	HG-AAS		FAAS		ICP-OES		ICP-MS		Spectrophotometry	
					n [a]	Mean ±SD [b]	*n*	Mean ± SD	*n*	Mean ±SD	*n*	Mean ±SD	*n*	Mean±SD
1 A	Pb	2.5633	3.0912	0.0823	- [c]	-	91	14.7 ± 0.5	5	14.5 ± 0.3	3	14.2± 0.4	-	-
	As	0.5863	3.0943	0.5584	89	3.25 ± 0.29	-	-	5	3.38 ± 0.34	-	-	3	3.18±0.18
1 B	Pb	1.7270	3.0922	0.1833	-	-	91	35.4 ± 1.3	5	34.3 ± 2.2	3	35.2± 1.1	-	-
	As	0.1426	3.0943	0.8673	89	8.24 ± 0.73	-	-	5	8.07 ± 0.27	-	-	3	8.22±0.13

2 A	Pb	0.2255	3.0766	0.7985	-	-	103	17.8 ± 1.2	6	17.8 ± 0.7	7	18.2 ± 1.2	-	-
	As	0.5656	2.6895	0.6388	92	3.60 ± 0.25	-	-	5	3.61 ± 0.56	6	3.66 ± 0.22	7	3.52 ± 0.24
2 B	Pb	0.5828	3.0766	0.5600	-	-	103	30.2 ± 1.4	6	29.8 ± 1.3	7	29.6 ± 1.5	-	-
	As	0.1069	2.6879	0.9559	92	7.22 ± 0.56	-	-	5	7.33 ± 1.41	-	-	10	7.18 ± 1.20

[a] *N*umber of the reported data after elimination of outliers by the Grubbs test,

[b] Units are mg kg^{-1} for mean value and SD,

[c] Not detection

Table 4

Results of robust statistics for the PT scheme [a]

Analyte	Pb				As			
	1 A(or ZB)	1 B(or ZW)	2 A(or ZB)	2 B(or ZW)	1 A(or ZB)	1 B(or ZW)	2 A(or ZB)	2 B(or ZW)
Number of results	99	99	116	116	99	99	115	115
Range of results	13.7-17.6	26.9-40.4	11.7-21.3	20.4-35.5	0.083-4.77	0.083-10.62	1.86-15.1	2.75-21.0
Assigned value (robust median)	15.2	35.4	18.0	30.2	3.24	8.32	3.60	7.16
Robust CV	2.32	2.25	5.77	2.95	6.06	4.37	3.29	3.73
Number (percentage) of results of $\lvert ZB \rvert$ or $\lvert ZW \rvert \leq 2$	82(82.8)	86(86.9)	103(88.8)	95(81.9)	82(82.8)	79(79.8)	99(86.1)	100(87.0)
Number (percentage) of results of $\lvert ZB \rvert$ or $\lvert ZW \rvert$ between 2 and 3	7(7.1)	4(4.0)	7(6.0)	8(6.9)	7(7.1)	5(5.2)	7(6.1)	7(6.1)
Number (percentage) of results of $\lvert ZB \rvert$ or $\lvert ZW \rvert \geq 3$	10(10.1)	9(9.1)	6(5.2)	13(11.2)	10(10.1)	15(15.2)	9(7.8)	8(7.0)

[a] Units are mg kg^{-1} for result and assigned value, % for robust CV and percentage

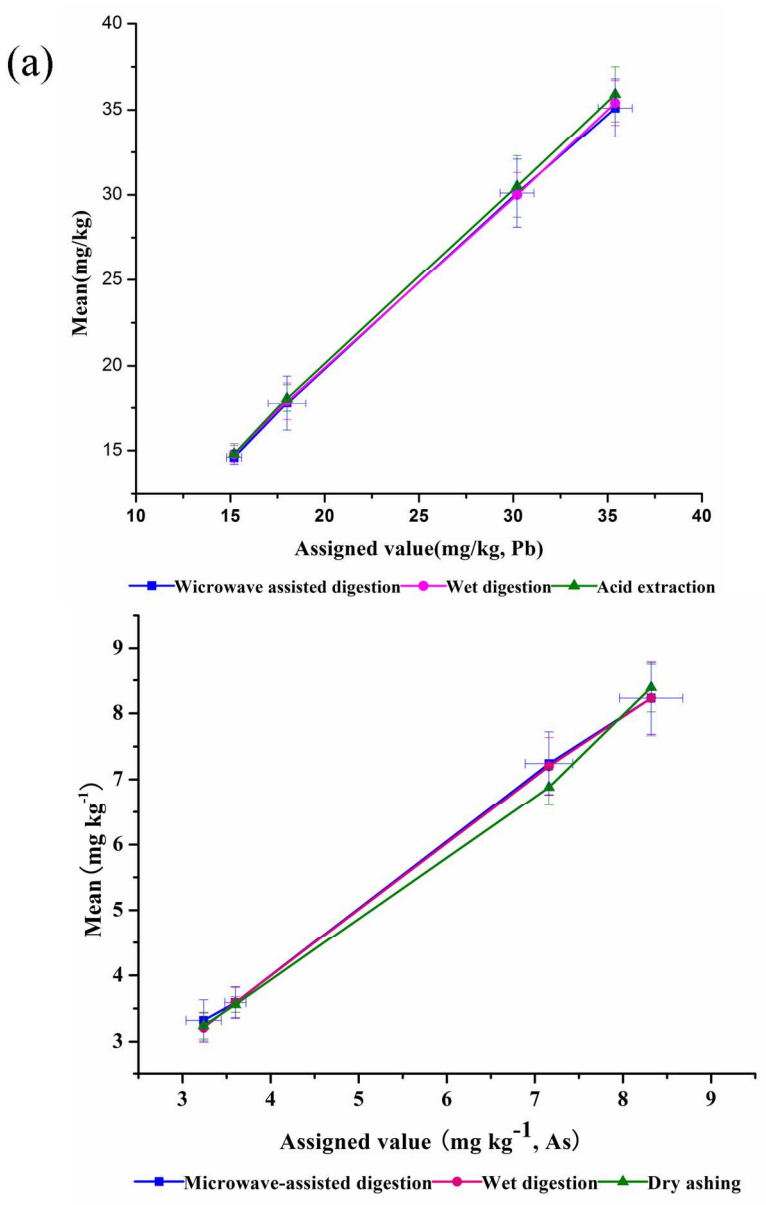


Fig. 1. Linear regression analyses of results obtained from different procedures. (a) Pb: microwave-assisted digestion; y=1.012x-0.59, r=0.9998; wet digestion, y=1.012x-0.47, r=0.9999; acid extraction, y=1.036x-0.77, r=0.9998; As: microwave-assisted digestion, y=0.985x+0.10, r=0.9996; wet digestion, y=0.995x+0.004, r=0.9998; dry ashing, y=0.991x-0.13, r=0.9982;
400x600mm (96 x 96 DPI)

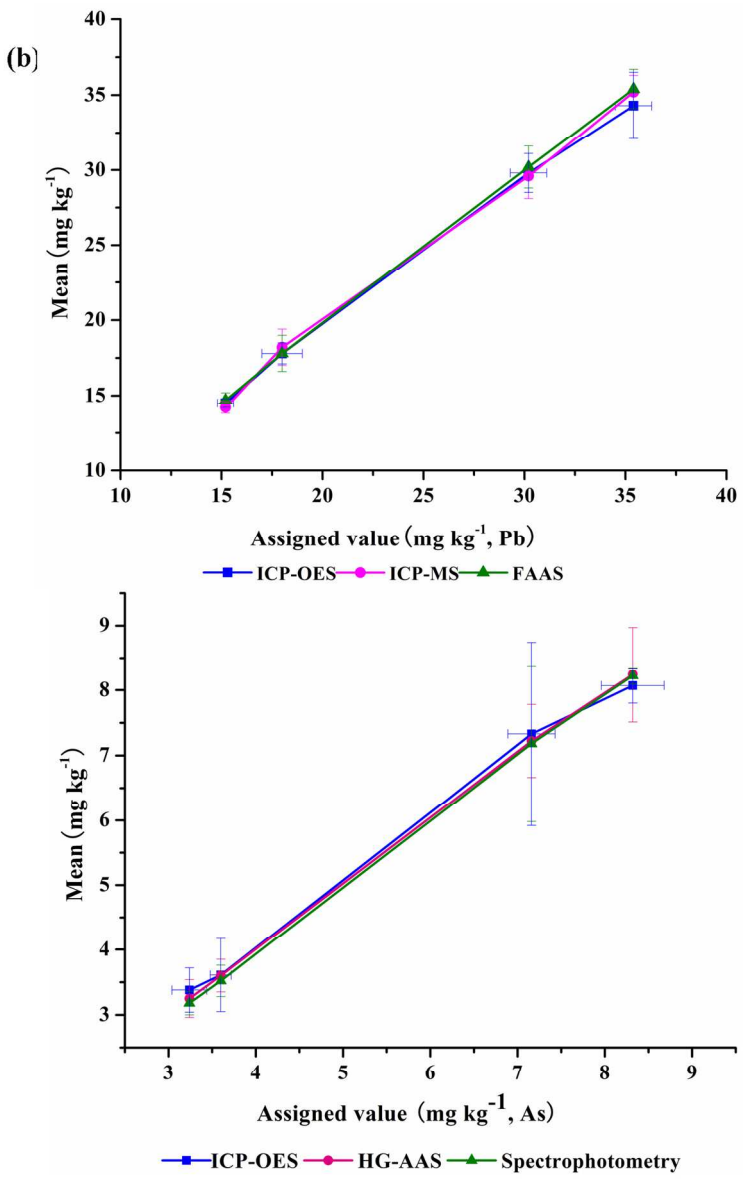


Fig. 1. Linear regression analyses of results obtained from different procedures.(b) Pb: ICP-OES, y=0.979x-0.07, r=0.9994; ICP-MS, y=1.010x-0.64, r=0.9986; FASS, y=1.022x-0.71, r=0.9999; As: ICP-OES, y=0.961x+0.24, r=0.9978; HG-AAS, y=0.992x+0.04, r=0.9997; spectrophotometry, y=1.003x-0.07, r=0.9998.
400x600mm (96 x 96 DPI)

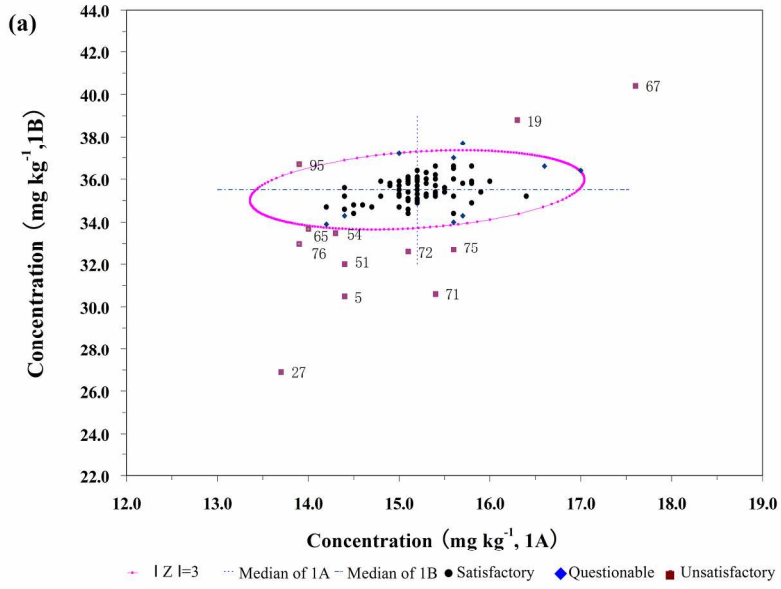


Fig. 2. Youden two-sample plots of concentrations for (a) Pb in samples 1 A and 1 B
209x148mm (300 x 300 DPI)

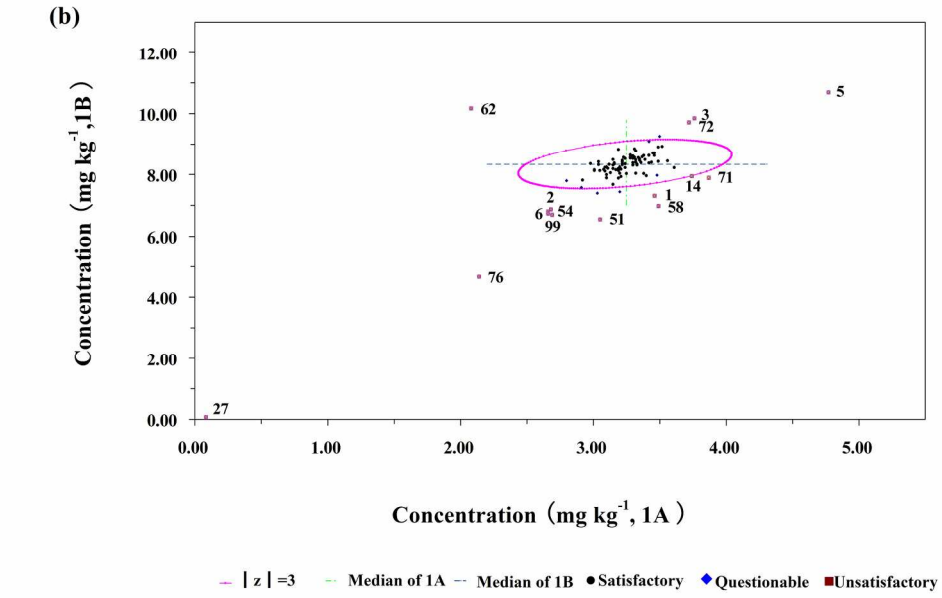


Fig. 2. Youden two-sample plots of concentrations for (b) As in samples 1 A and 1 B, 209x148mm (300 x 300 DPI)

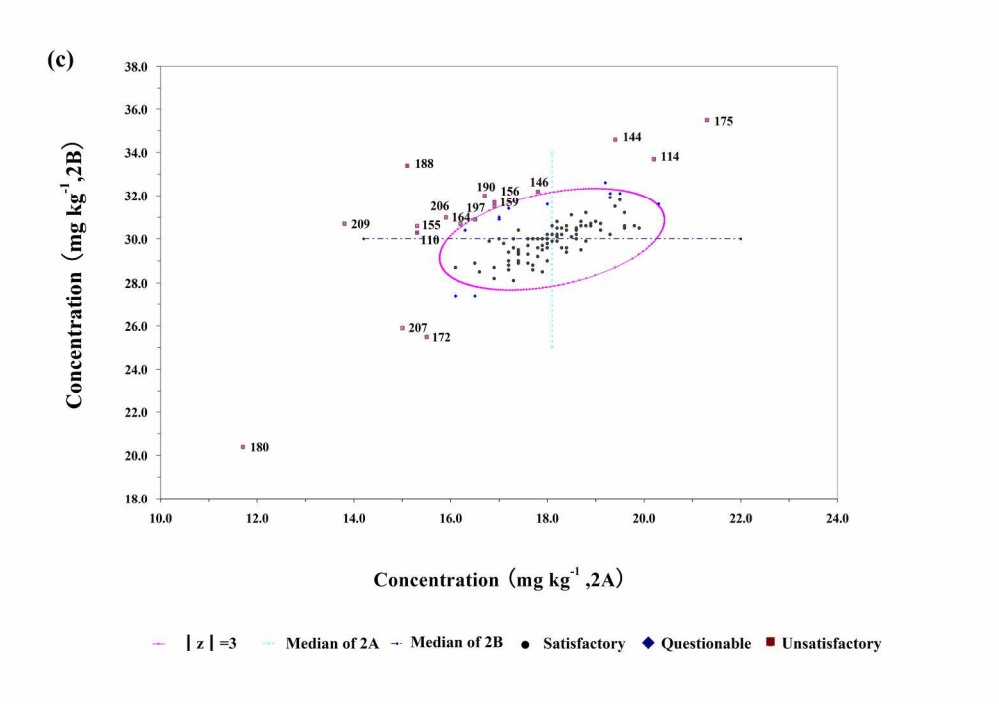


Fig. 2. Youden two-sample plots of concentrations for (c) Pb in samples 2 A and 2 B, 209x148mm (300 x 300 DPI)

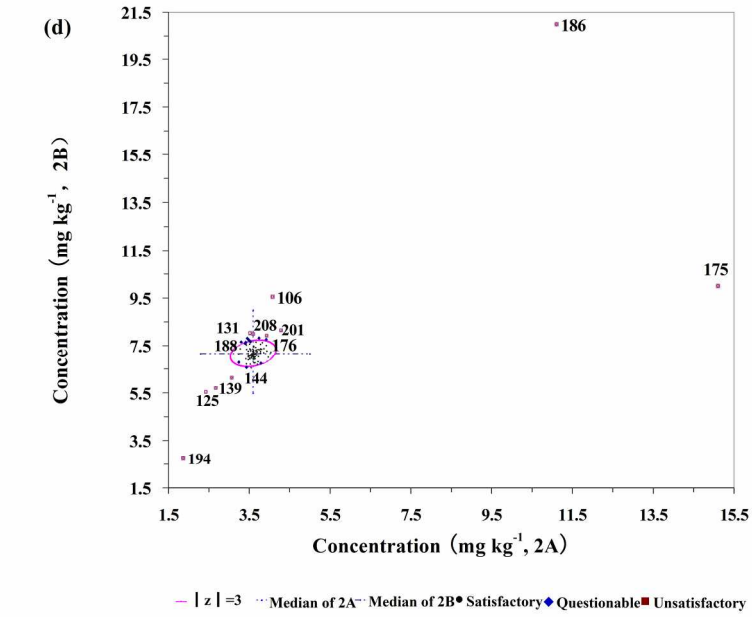


Fig. 2. Youden two-sample plots of concentrations for (d) As in samples 2 A and 2 B. The X-axis and Y-axis indicated the average concentrations (mg kg-1) of Pb or As in A and B samples. The critical points (=3) were plotted as ellipse at a confidence level of 99%.
209x148mm (300 x 300 DPI)

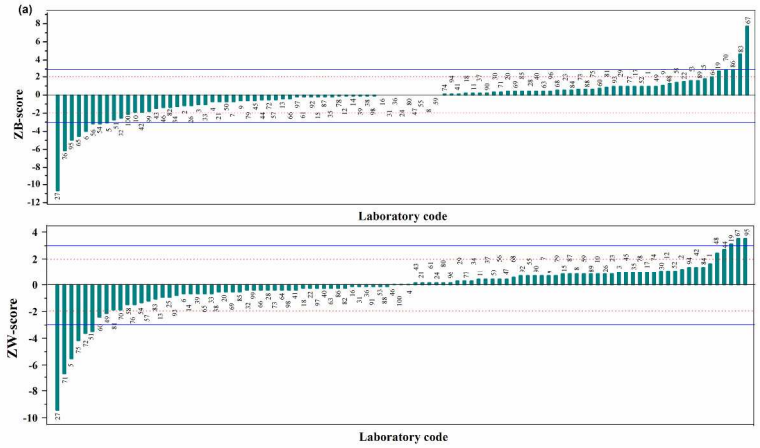


Fig.3. Participants' ZB- and ZW-score values for (a) Pb in samples 1 A and 1 B
288x202mm (300 x 300 DPI)

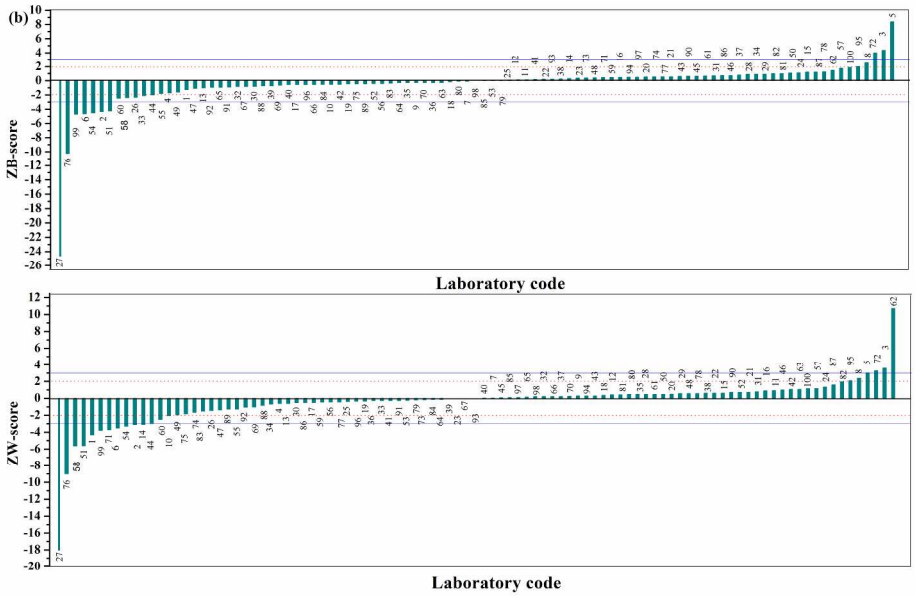


Fig.3. Participants' ZB- and ZW-score values for (b) As in samples 1 A and 1 B
296x209mm (300 x 300 DPI)

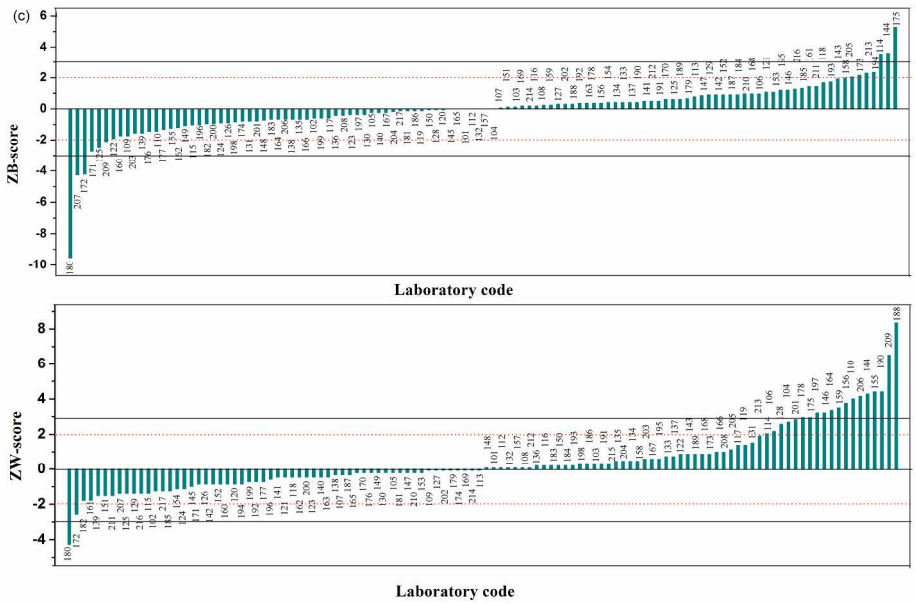


Fig.3. Participants' ZB- and ZW-score values for(c) Pb in samples 2 A and 2 B
296x209mm (300 x 300 DPI)

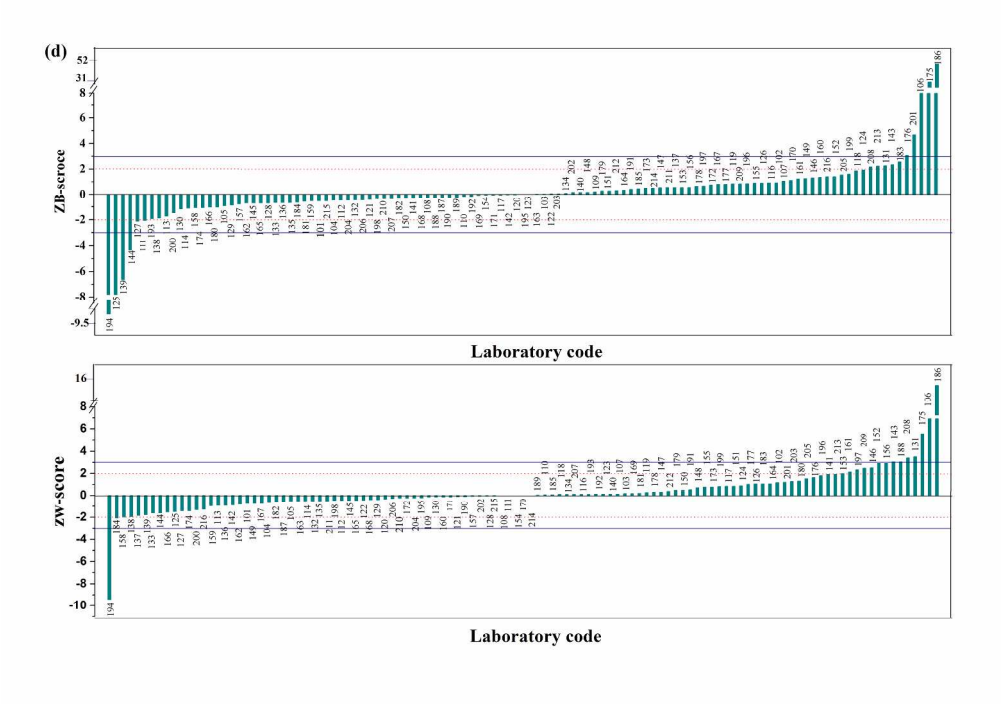


Fig.3. Participants' ZB- and ZW-score values for (d) As in samples 2 A and 2 B. Dotted and solid lines correspond to =2 and =3
296x209mm (300 x 300 DPI)