Talanta 130 (2014) 462–469

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Monte Carlo simulation of expert judgments on human errors in chemical analysis—A case study of ICP–MS

Ilya Kuselman [a,*], Francesca Pennecchi [b], Malka Epstein [a], Ales Fajgelj [c], Stephen L.R. Ellison [d]

[a] National Physical Laboratory of Israel (INPL), Danciger "A" Bldg, Givat Ram, 91904 Jerusalem, Israel
[b] Istituto Nazionale di Ricerca Metrologica (INRIM), 91 Strada delle Cacce, 10135 Turin, Italy
[c] International Atomic Energy Agency (IAEA), Vienna International Centre, PO Box 100, A-1400 Vienna, Austria
[d] Laboratory of Government Chemist Ltd (LGC), Queens Road, Teddington TW11 0LY, Middlesex, UK

ARTICLE INFO

Article history:
Received 11 May 2014
Received in revised form
9 July 2014
Accepted 11 July 2014
Available online 19 July 2014

Keywords:
Monte Carlo simulation
Expert judgment
Human error
Chemical analysis
ICP–MS

ABSTRACT

Monte Carlo simulation of expert judgments on human errors in a chemical analysis was used for determination of distributions of the error quantification scores (scores of likelihood and severity, and scores of effectiveness of a laboratory quality system in prevention of the errors). The simulation was based on modeling of an expert behavior: confident, reasonably doubting and irresolute expert judgments were taken into account by means of different probability mass functions (pmfs). As a case study, 36 scenarios of human errors which may occur in elemental analysis of geological samples by ICP–MS were examined. Characteristics of the score distributions for three pmfs of an expert behavior were compared. Variability of the scores, as standard deviation of the simulated score values from the distribution mean, was used for assessment of the score robustness. A range of the score values, calculated directly from elicited data and simulated by a Monte Carlo method for different pmfs, was also discussed from the robustness point of view. It was shown that robustness of the scores, obtained in the case study, can be assessed as satisfactory for the quality risk management and improvement of a laboratory quality system against human errors.

1. Introduction

Quality risk management in a chemical analytical laboratory in the pharmaceutical industry, medicine or any other field requires identification and mapping of human errors as potential hazards which may occur during the analysis (measurement/testing process). Evaluation of the error likelihood and severity (the risk assessment) is necessary for their reduction, i.e., mitigation, avoidance or blocking of the errors by the laboratory quality system [1,2].

There are commission errors (knowledge-, rule- and skill-based mistakes and routine, reasoned, reckless and malicious violations) and omission errors (lapses and slips) of a sampling inspector and/or an analyst/operator [3]. These errors are active. Errors due to a poor laboratory design, a defect in the equipment or a faulty management decision, not depending on the inspector and/or the operator, are defined as latent errors [2]. The kinds of human error $k=1, 2, \ldots, K$ and steps of the analysis $m=1, 2, \ldots, M$ in which the error may happen (locations of the error), form event scenarios $i=1, 2, \ldots, I$, where $I=K\times M$.

A Swiss cheese model is used for characterization of the interaction of errors with a laboratory quality system [3]. This model considers the quality system components $j=1, 2,\ldots, J$ as protective layers against human errors. For example, the system components are: validation of the measurement/analytical method and formulation of standard operation procedures (further "validation"); training of analysts and proficiency testing (further "training"); quality control using statistical charts and/or other means (further "quality control"); and supervision. Each such component has weak points, whereby errors are not prevented, similar to holes in slices of the cheese. Coincidence of the holes in all components of the laboratory quality system on the path of a human error is a defect of the quality system, permeable to the error, leading to an atypical result of the analysis.

By a recently developed technique for error quantification [4], an expert in the analysis may estimate likelihood p_i of scenario i by the following scale: likelihood of an unfeasible scenario—as $p_i=0$, weak likelihood—as $p_i=1$, medium—as $p_i=3$, and strong (maximal) likelihood—as $p_i=9$. The expert estimates/judgments on severity of an error by scenario i as expected loss l_i of reliability of the analysis, are performed with the same scale (0, 1, 3, 9). Estimates of the possible reduction r_{ij} of likelihood and severity of human error scenario i as a result of the error blocking by quality system layer j (degree of interaction) are made by the expert

* Corresponding author. Tel.: +972 2 6303501; fax: +972 2 6303516.
E-mail address: ilya.kuselman@gmail.com (I. Kuselman).

http://dx.doi.org/10.1016/j.talanta.2014.07.036
0039-9140/

(s) using the same scale again. The interrelationship matrix of r_{ij} has I rows and J columns, hence it contains $I \times J$ entries. Blocking human error according to scenario i by a quality system component j can be more effective in the presence of another component j' ($j' \neq j$) because of their synergy $\Delta_{jj'}^{(i)}$, equals to 0 when the effect is absent, and equals to 1 when it exists. Estimates q_j of importance of quality system component j in decreasing losses from human error are calculated as $q_j = \sum_{i=1}^{I} p_i l_i r_{ij} s_{ij}$, where the synergy factor is $s_{ij} = 1 + \sum_{j' \neq j}^{J} \Delta_{jj'}^{(i)} / (J-1)$.

The technique allows transformation of the independent semi-intuitive expert judgments on human errors and on the laboratory quality system into the following quantitative scores expressed in %:

(a) likelihood score of human error in the analysis $P^* = (100\%/9)\sum_{i=1}^{I} p_i / I$;
(b) severity (loss) score of human error for reliability of the analysis results $L^* = (100\%/9)\sum_{i=1}^{I} l_i / I$;
(c) importance score of a component of the laboratory quality system $q_j^* = 100\%\, q_j / \sum_{j=1}^{J} q_j$, and similar score of the quality system influence at step m of the analysis $\tilde{q}_m^* = 100\%\, \tilde{q}_m / \sum_{m=1}^{M} \tilde{q}_m$, where $\tilde{q}_m = \sum_{i'=m}^{m+M(K-1)} \sum_{j=1}^{J} p_{i'} l_{i'} r_{i'j} s_{i'j}$, and $i' = m + M(k-1)$ are the scenario numbers related to the same error location (step m) for all kinds of error $k = 1, 2, \ldots, K$; and
(d) effectiveness score of the quality system, as a whole, against human error $Eff^* = (100\%/9)\sum_{j=1}^{J} q_j / \sum_{j=1}^{J} \sum_{i=1}^{I} p_i l_i s_{ij}$.

This technique was applied for quantification of human errors in pH measurements of groundwater [4], and in multi-residue analysis of pesticides in fruits and vegetables [5]. Expert judgments are used also in landscape ecology [6] and biosecurity [7], counterterrorism risk assessment [8,9] and many other fields [10]. A comprehensive list of expert judgment bibliography is available in Ref. [11]. In analytical chemistry and metrology an expert judgment is based on knowledge of the nature of the analyte and measurand, the analytical procedure (measurement method) used, earlier observations and common sense. Thus, judgments are not arbitrary [12]. However, any expert is also a human being and the elicitation process [13], by which the expert is prompted to provide error likelihood, severity and other estimates, is influenced by aleatory and epistemic uncertainty [14], intrapersonal conflicts [15], etc.

Therefore, evaluation of variability of the error quantification scores (their robustness to the doubts of an expert) is important as well [5]. The present paper describes such an evaluation based on Monte Carlo simulations. Scenarios of human errors in elemental analysis of geological samples (determination of elemental mass fractions) by inductively coupled plasma mass spectrometry (ICP–MS) and variability of the corresponding error quantification scores are discussed here as a case study.

2. An expert judgment as a discrete quantity

An expert judgment for human error quantification is a discrete quantity that can take any scale value among (0, 1, 3, 9) with a probability distributed according to the judgment probability mass function (pmf). When a value is chosen on the scale, the expert may still feel a doubt concerning neighboring scale values. Choosing 0, this expert thinks about 1 as a value which is possible, but with significantly lower pmf. Choosing 1, the expert necessarily takes into account 0 and 3, but again with equally lower pmf, etc. However, more distant scale values are not relevant. Otherwise, the expert is not experienced in the field and should not participate in the elicitation process. The following distributions modeling an expert behavior are studied in the present paper:

1) of confident expert judgments: the pmf at a chosen scale value is 0.90, whereas close values on the right and/or on the left on the scale have a total pmf equal to 0.10;
2) of reasonably doubting expert judgments: the pmf at a chosen value is 0.70, whereas close values on the scale have a total pmf equal to 0.30; and
3) of irresolute expert judgments: the pmf at a chosen value is 0.50, and the close values on the scale have a total pmf equal to 0.50 also.

Table 1
Probability mass functions (pmfs) of expert judgments.

Expert judgment	Chosen scale value	Scale			
		0	1	3	9
Confident	0	0.90	0.10	0.00	0.00
	1	0.05	0.90	0.05	0.00
	3	0.00	0.05	0.90	0.05
	9	0.00	0.00	0.10	0.90
Reasonably doubting	0	0.70	0.30	0.00	0.00
	1	0.15	0.70	0.15	0.00
	3	0.00	0.15	0.70	0.15
	9	0.00	0.00	0.30	0.70
Irresolute	0	0.50	0.50	0.00	0.00
	1	0.25	0.50	0.25	0.00
	3	0.00	0.25	0.50	0.25
	9	0.00	0.00	0.50	0.50

More pmf details are shown in Table 1. These three pmfs represent properly the whole range of cases from the most to the less confident expert judgments in the framework of the proposed modeling.

Sampling from these distributions for random generation of expert judgments as discrete values was performed using a code developed in *R* [16,17].

3. Distributions of score values for quantification of human errors

Since human error quantification scores P^*, L^*, q_j^*, $\tilde{q}_m^*$ and Eff^*, are calculated as algebraic combinations of the elicited expert judgments p_i, l_i, r_{ij}, and synergy factors s_{ij} (which, in the present context, are considered as entirely known), the probability distributions for these scores depend on the distributions of the expert judgments. A Monte Carlo simulation of the score distributions was performed based on the following algorithm inspired by JCGM 101 [18]:

1) input of the elicited estimates p_i, l_i and r_{ij}, synergy factors s_{ij}, numbers K of kinds of human error, M of steps of the chemical analysis, I of human error scenarios, J of the laboratory quality system components, and number of the Monte Carlo trials $n_{MC} = 100{,}000$;
2) assignment of probability mass functions (pmfs) to the expert judgments p_i, l_i, r_{ij};
3) simulation of possible values of expert judgments on human error by scenario i according to the chosen pmf on the scale values (0, 1, 3, 9) for $i = 1$ to I: the matrix of simulated values is of dimension $I \times n_{MC}$;
4) determination of the numerical distributions for scores P^*, L^*, q_j^*, $\tilde{q}_m^*$ and Eff^* by propagating the simulated distributions of the expert judgments into the relevant equations discussed in

the Introduction; evaluation of the score mean, median and standard deviation (mean and median can be different because of a possible asymmetry in the simulated distributions);

5) plotting histograms for the distributions of the scores.

The detailed *R* code can be sent upon request.

4. Steps of elemental analysis of geological samples by ICP–MS and some analytical details

There are four main steps $m=1$, 2, 3, and 4 of the analysis: (1) sample preparation, (2) ICP–MS calibration; (3) measurement with ICP–MS; and (4) calculation of elemental mass fractions in analyzed samples and reporting.

Sample preparation of rocks and sediments is based on fusion of the sample with lithium metaborate or sodium peroxide flux. Then the obtained bead is dissolved in nitric acid in an ultrasonic bath [19,20]. The solution should be filtered and diluted with water to a sample/solution weight ratio in the range from 1:1000 to 1:4000. Samples of peridotites and a number of types of magma can be prepared by digestion of a sample with an HF–HNO_3 mixture in an ultrasonic bath. When samples contain resistant phases, e.g. zircon, applied temperature and pressure are increased using microwaves or digestion bombs. Then samples are evaporated to incipient dryness, refluxed in nitric acid, evaporated and dissolved again, filtered and diluted with water [21]. For analysis of trace and rare earth elements sample digestion with an HF–$HClO_4$ mixture under pressure can be applied [22]. In any case an analytical blank is prepared identically to the samples.

Synthetic and natural certified reference materials (CRMs) [23] are used for preparation of matrix matched calibrators of ICP–MS [24]. Concentration of the acids and flux quantity in such calibrators should be the same as in the samples prepared for analysis. That is in addition to the known requirement to CRMs used to have a composition close to the analyzed samples for minimization of matrix effects [25]. The CRMs (not the same as for calibration) are used also as internal standards and quality control samples [26].

5. Scenarios of human error

5.1. Knowledge-based mistakes, $k=1$

A knowledge-based mistake occurs when an analyst faces a new situation, wherein his/her knowledge for making the right decision is not sufficient.

Scenario $i=1$ in sample preparation, $m=1$. A sample containing an excessively high quantity of an analyte (not diluted as necessary) may produce too low of a response since not all the quantity will be ionized, resulting in a wrong recovery factor applied for corrections.

Scenario $i=2$ in ICP–MS calibration, $m=2$. Application of an inadequate calibrator (with a difference in the matrix in comparison to the samples) may lead to a bias in the test results.

Scenario $i=3$ in measurement with ICP–MS, $m=3$. Use of an improper blank solution (did not pass all the steps of the sample preparation) may also cause biased results.

Scenario $i=4$ in calculation and reporting, $m=4$. Mistaken interpretation of interferences (e.g., due to diatomic molecules) may influence the test result.

5.2. Rule-based mistakes, $k=2$

A rule-based mistake may occur when an analyst encounters a relatively familiar problem, but applies an unsuitable solution or rule.

Scenario $i=5$ in sample preparation, $m=1$. An analyst, using as a rule sample preparation by digestion of a sample with an HF–HNO_3 mixture in an ultrasonic bath, may not take into account that a sample contains a resistant phase, which requires application of a microwave or digestion bombs.

Scenario $i=6$ in ICP–MS calibration, $m=2$. Usual dilution of reference materials for preparation of calibrators, when another dilution is necessary, may cause atypical test results.

Scenario $i=7$ in measurement with ICP–MS, $m=3$. When drift of the instrument response is usually controlled for specific ion masses, whereas another analyte is under determination, the control may be not sufficient and the results shifted.

Scenario $i=8$ in calculation and reporting, $m=4$. Unusual sample mass applied in an analysis (e.g., to increase quantity of an analyte) may be forgotten by an operator and the regular mass introduced erroneously in the file for calculations.

5.3. Skill-based mistakes, $k=3$

Skill-based mistakes are the result of inadequate analyst performance occurring from overconfidence of the type “I have done this many times” [27].

Scenario $i=9$ in sample preparation, $m=1$. Dissolution of a sample in an acid mixture containing HF in a Teflon beaker (not in a digestion bomb) as usually done for determination of minor elements and/or traces, wherein silicon is an analyte, may lead to a loss of silicon.

Scenario $i=10$ in ICP–MS calibration, $m=2$. Use of the same calibrator as previously, when its container is not closed hermetically and the element concentrations change due to water evaporation, is a mistake.

Scenario $i=11$ in measurement with ICP–MS, $m=3$. Flux-fusion sample solutions may form a gel, not always immediately visible, but clogging the nebulizer and leading to inhomogeneity of the analyte distribution in the test portion. Measurements of the analyte concentrations in such solutions (in regular conditions) may lead to mistaken results.

Scenario $i=12$ in calculation and reporting, $m=4$. A skill-based human error is possible when an analyst usually uses a certain order of samples, whereas an assisting operator arranged the samples in another way.

5.4. Routine violations, $k=4$

As a rule the reason for routine violations in the analysis is a wish to shorten the work.

Scenario $i=13$ in sample preparation, $m=1$. A decision of an analyst after dissolution (based on visual inspection) that the filtration is not necessary and may be ignored to shorten the procedure, is a routine violation.

Scenario $i=14$ in ICP–MS calibration, $m=2$. To prepare a calibrator, a small value of concentrated CRM solution may be diluted to a large volume by one step, to avoid spending time for a longer procedure with more steps of dilution. The calibrator prepared in this way will be not accurate.

Another example is reducing the number of calibrators in order to shorten the work.

Scenario $i=15$ in measurement with ICP–MS, $m=3$. Reducing the required number of replicates for shortening the work is a routine violation.

Another example is when result reading is started immediately after introduction of a sample into the instrument, without waiting at least 1 min for a stable response.

Scenario $i=16$ in calculation and reporting, $m=4$. When a report is not checked with purpose to save time, a twist of the data during their transformation to the final file may not be noted,

and therefore not corrected, as in any other determination of a number of analytes in a number of samples, e.g., in Ref. [5].

5.5. Reasoned violations, $k=5$

A reasoned violation is caused by a wish to improve the analytical process.

Scenario $i=17$ in sample preparation, $m=1$. An analyst may use more flux for fusion than required by the procedure in order to improve a sample preparation. However, it will lead to an increased concentration of salts, not appropriate for the blank in the run (for a set of samples).

Scenario $i=18$ in ICP–MS calibration, $m=2$. To improve a method, an analyst may wish to increase a calibration range (which is anyway wide in ICP–MS) in spite of limitation at both minimal and maximal analyte concentrations.

Scenario $i=19$ in measurement with ICP–MS, $m=3$. When a flow-injection system is used for the sample introduction, a limited number of analyte concentrations can be measured simultaneously (in the same run). An attempt to increase the number of analytes is a routine violation, since some of the analytes may not be detected accurately as required.

Scenario $i=20$ in calculation and reporting, $m=4$. An analyst may remove an outlier from the data without investigation with the purpose to obtain a more "accurate" test result.

Another example is the "reference materials syndrome", when an analyst reports analyte concentration values close to those in CRM certificates (applied as control samples) which are subsequently found to be incorrect [28].

5.6. Reckless violations, $k=6$

A reckless violation may be a result of a state of mind in which an analyst acts without caring about the consequences.

Scenario $i=21$ in sample preparation, $m=1$. When cleaning of crucibles for fusing or glassware for dilution is performed improperly, a sample may become contaminated.

Scenario $i=22$ in ICP–MS calibration, $m=2$. Use of a CRM after the expiration date may lead to a biased calibration curve.

Scenario $i=23$ in measurement with ICP–MS, $m=3$. If an inadequate blank (from a previous analysis run) is taken recklessly, the measurement results may be biased.

Another example is when an analyst does not notice that a blank may also produce a response caused or influenced by contamination.

Scenario $i=24$ in calculation and reporting, $m=4$. Recklessness may lead to confusing names of samples.

5.7. Malicious violations, $k=7$

A malicious violation, including sabotage, is possible as a result of a conflict between an analyst and the laboratory manager [29].

Scenario $i=25$ in sample preparation, $m=1$. It may be reflected in filling sample labels in a confusing manner, i.e., such labels cannot be read simply and unambiguously.

Scenario $i=26$ in ICP–MS calibration, $m=2$. On this step of the analysis the violation may consist of the use of previous calibration data instead of re-calibration.

Scenario $i=27$ in measurement with ICP–MS, $m=3$. Confusing names of samples may be caused not only because of recklessness, as in scenario 24 above, but also intentionally.

Scenario $i=28$ in calculation and reporting, $m=4$. Falsification of data is a malicious violation.

5.8. Lapses, $k=8$

A lapse is an occurrence in which an analyst fails to act as required for a brief time (e.g., "senior moments" [30]).

Scenario $i=29$ in sample preparation, $m=1$. An analyst may forget that he/she has already added the necessary quantity of an internal standard to a sample and adds it again.

Another example is when the analyst forgets to dry a sample before weighing.

Scenario $i=30$ in ICP–MS calibration, $m=2$. An analyst/operator may forget to stir a prepared calibrator as required.

Scenario $i=31$ in measurement with ICP–MS, $m=3$. Cleaning of a nebulizer and/or glassware used between runs may be forgotten because of a lapse.

Scenario $i=32$ in calculation and reporting, $m=4$. A lapse may happen during the introduction of the data into the file: incorrect test results may be reported in such a case.

5.9. Slips, $k=9$

A slip is an action of an analyst that is not in accordance with the plan.

Scenario $i=33$ in sample preparation, $m=1$. A sample may be incompletely transferred into a crucible, when poured out by a slip after weighing.

Scenario $i=34$ in ICP–MS calibration, $m=2$. When an analyst pushes the arm of an automatic pipette stronger than necessary to achieve the stop (because of a slip), the volume is larger than required and the concentration of the analyte in the calibrator is changed.

Scenario $i=35$ in measurement with ICP–MS, $m=3$. If a capillary used for sample introduction is set inaccurately by a slip, and air is passed with the liquid to the nebulizer, the measurement results may be erroneous.

Scenario $i=36$ in calculation and reporting, $m=4$. When the daily number of analyzed samples (sample throughput) is large, reporting results related to one sample as results of another sample, by a slip, is possible.

Thus, in spite of achievements in ICP–MS development, there are still a number of human error scenarios in elemental analysis of geological samples, which should be taken into account in a routine laboratory for quality risk management.

6. Results and discussion

Elicited expert judgments on human errors by the described 36 scenarios are presented in Table 2. Results of the direct score calculations from these data (using formulas discussed in the Introduction) are in Table 3. Mean, median and standard deviation (STD) of the relevant distributions determined using Monte Carlo simulations are also presented in Table 3. One can see from Table 3 that the mean score values of a confident expert are very close to the score values calculated directly from the elicited data. The mean values (as well as the median) change systematically depending on the confidence of the expert judgments, i.e., depending on the relevant pmfs. Accordingly, it makes sense that the corresponding standard deviations increase as long as the expert judgments become less confident. However, it is interesting that all the mean (and the median) values of the simulated scores remain consistent with the score values calculated directly from the data within two such standard deviations. Moreover, the score values calculated directly from the data can be interpreted as obtained when the expert judgments are "absolutely confident", i.e., when a Dirac delta function, centered at a specific expert

estimate on the scale, is applied as the pmf. This pmf takes value 1 at that expert estimate (judgment), whereas the same pmf equals to zero at the rest of the scale.

Histograms illustrating the score distributions for a reasonably doubting expert, for example, are shown in Fig. 1. There are: (a) the likelihood score P^*, (b) the severity score L^*, (c) the importance of quality control score $q^*_{j=3}$, (d) the score $\tilde{q}^*_{m=2}$ of the quality system influence at ICP–MS calibration (second step of the analysis), and (e) the effectiveness score Eff^* of the quality system in the analysis in whole.

Table 2
The elicited expert judgments.

Scenario i	Likelihood p_i	Severity l_i	Degree of interaction r_{ij} Quality system component j				Synergy factor s_{ij} Quality system component j			
			1	2	3	4	1	2	3	4
1	3	9	3	9	1	3	1.67	1.33	1.33	1
2	1	3	3	3	3	9	1.67	1.33	1.33	1
3	3	3	3	9	9	9	1.67	1.33	1.33	1
4	1	1	9	9	3	9	1.67	1.33	1.33	1
5	3	3	9	3	3	3	1.67	1.33	1.33	1
6	1	1	3	3	9	3	1.67	1.33	1.33	1
7	3	3	1	9	9	3	1.67	1.33	1.33	1
8	3	9	1	3	3	3	1.67	1.33	1.33	1
9	1	3	3	3	3	3	1.67	1.33	1.33	1
10	3	1	1	3	3	3	1.67	1.33	1.33	1
11	3	9	3	3	9	3	1.67	1.33	1.33	1
12	3	9	3	3	1	1	1.67	1.33	1.33	1
13	3	3	3	3	9	3	1.67	1.33	1.33	1
14	3	3	9	9	3	9	1.67	1.33	1.33	1
15	3	3	1	3	9	9	1.67	1.33	1.33	1
16	3	9	3	3	3	3	1.67	1.33	1.33	1
17	1	1	9	9	3	3	1.67	1.33	1.33	1
18	1	3	9	9	3	3	1.67	1.33	1.33	1
19	3	3	3	9	3	3	1.67	1.33	1.33	1
20	3	3	3	3	3	3	1.67	1.33	1.33	1
21	1	3	3	3	9	3	1.67	1.33	1.33	1
22	3	3	3	3	3	3	1.67	1.33	1.33	1
23	1	3	1	9	3	3	1.67	1.33	1.33	1
24	3	9	0	3	3	3	1.67	1.33	1.33	1
25	1	9	0	1	1	1	1	1	1	1
26	1	9	0	1	3	3	1	1	1	1
27	1	9	0	1	3	3	1	1	1	1
28	1	9	0	1	3	3	1	1	1	1
29	3	3	0	1	1	1	1	1	1	1
30	3	3	0	3	3	3	1	1	1	1
31	1	9	1	3	3	3	1	1	1	1
32	1	9	0	1	3	3	1	1	1	1
33	1	3	0	1	1	1	1	1	1	1
34	1	3	0	3	3	3	1	1	1	1
35	3	3	0	3	3	3	1	1	1	1
36	1	9	0	3	3	3	1	1	1	1

6.1. Likelihood and severity

The likelihood score, summarizing the elicited judgments presented in Table 2, is $P^*=22\%$. This means that human error may happen in the analysis on average of about one out of five samples. It is clear from Table 3 that a less confident expert may lead to larger estimates for P^*, from 24% to 29%, on average. Hence, the less confident the expert is, the more underestimated is the P^* value actually calculated from the data. When P^* is increasing, the standard deviation is also increasing from 2% to 4%. Similar P^* variability was also discussed in Ref. [5], based on the assumption that an expert may change his/her opinion "tomorrow" on 1 or 2 scenarios. In spite of equivalence of P^* mean and median (rounded) values for a reasonably doubting expert, a minor asymmetry of the histogram in Fig. 1a is visible, probably due to the non-equidistant scale of the expert estimates/judgments through which the input pmfs were propagated. However, there are also other possible reasons for the asymmetry, e.g., when an expert unconsciously avoids one of the extreme choices on the scale (0 and 9). The human error severity score is $L^*=56\%$. Thus, approximately every second result of the ICP–MS elemental analysis burdened with human errors cannot be corrected, and analysis of corresponding samples should be repeated, when the error source is detected and eliminated.

For comparison, in the example provided in Ref. [4] for pH measurement of groundwater, the likelihood score value (27%) and the severity score value (65%) were larger than in the current case. For multi-residue analysis of pesticides in fruits and vegetables $P^*=19\%$ and $L^*=84\%$ were obtained [5]. The difference between these score values can be explained by specificity of the analytical methods and conditions of their use in a routine laboratory.

As appears from Table 3, a less confident expert leads to a reduction in the estimates of the severity from mean $L^*=55\%$ to 51%, but again with an increasing standard deviation from 3% to 5%. There is no difference between mean and median for a reasonably doubting expert, but a minor asymmetry of the histogram in Fig. 1b is also observed, as for the likelihood score.

Table 3
The score values (%) calculated directly from the elicited data in comparison to those obtained by Monte Carlo simulations.

Score	Calculated directly	Monte Carlo simulations								
		Confident expert			Reasonably doubting			Irresolute expert		
		Mean	Median	STD	Mea	Median	STD	Mean	Median	STD
P^*	22	24	23	2	26	26	3	29	29	4
L^*	56	55	55	3	53	53	5	51	51	5
$q^*_{j=1}$	24	25	24	2	26	25	4	27	26	5
$q^*_{j=2}$	26	26	26	2	26	26	3	26	26	4
$q^*_{j=3}$	27	27	27	2	26	26	3	26	26	4
$q^*_{j=4}$	23	23	23	2	22	22	3	22	21	4
$\tilde{q}^*_{m=1}$	30	29	28	6	26	25	9	25	23	10
$\tilde{q}^*_{m=2}$	14	15	14	4	16	15	6	18	17	8
$\tilde{q}^*_{m=3}$	25	25	25	5	26	25	9	27	26	11
$\tilde{q}^*_{m=4}$	32	32	31	6	31	30	9	30	29	11
Eff^*	55	54	54	3	53	53	4	51	51	5

Note: STD is standard deviation of a simulated score value from its mean.

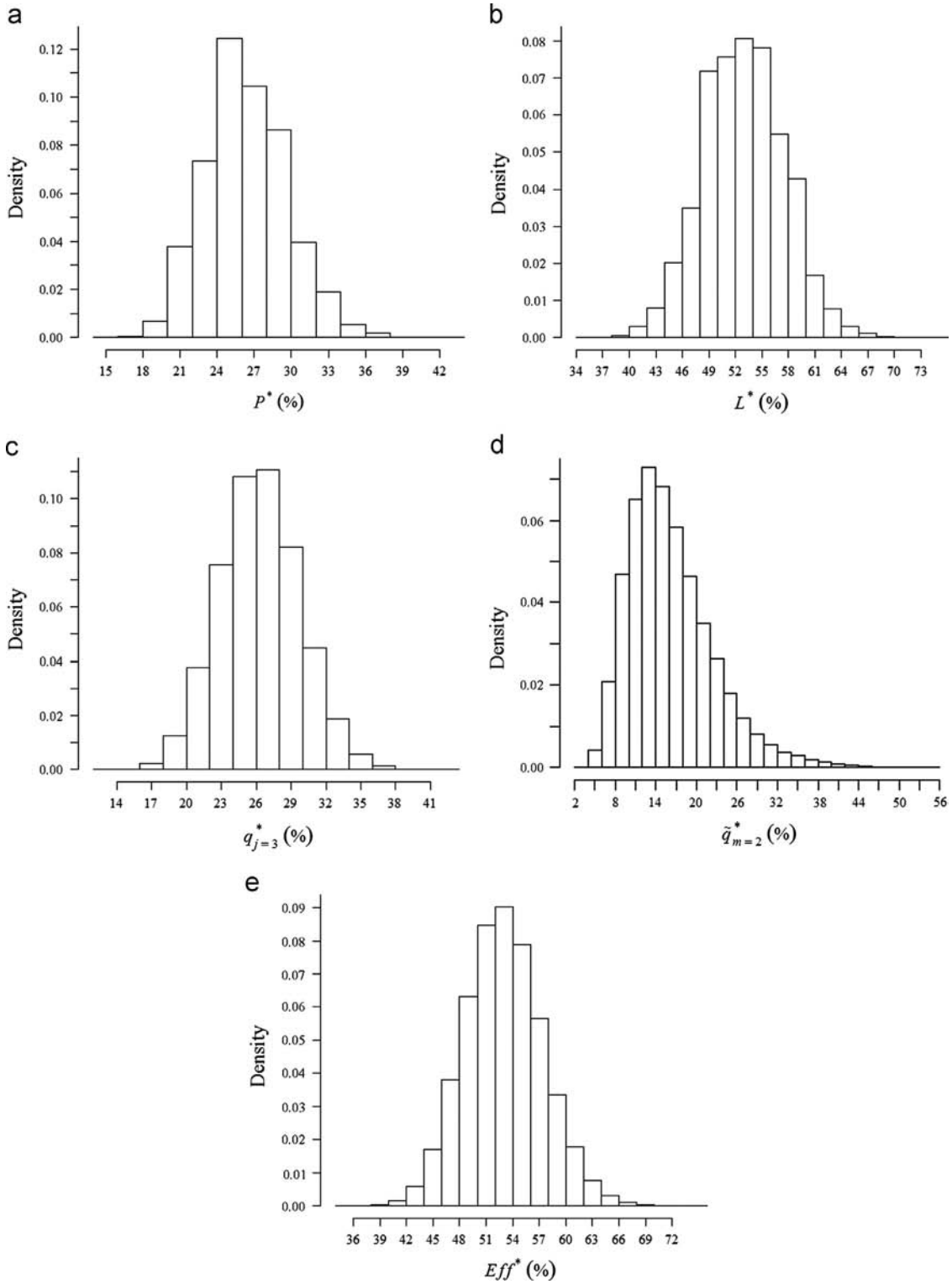

Fig. 1. Histograms of the scores corresponding to judgments of a reasonably doubting expert, simulated by the Monte Carlo method: (a) the likelihood P^*, (b) the severity L^*, (c) the importance of quality control $q^*_{j=3}$, (d) the influence of the quality system at second step of the analysis (ICP–MS calibration) $\tilde{q}^*_{m=2}$, and (e) the effectiveness of the quality system at all steps of the analysis in whole Eff^*.

6.2. Quality system scores

From Table 3 one can understand that the most important component of the quality system is quality control ($q^*_{j=3}$ is maximal), then training, validation and supervision follow. This is a different ordering than in Ref. [4,5]. The $q^*_{j=3}$ score value is from 27% to 26% for all discussed kinds of calculation and simulations, with a standard deviation varying from 2% to 4%. Neither a difference between the mean and the median of the simulated values in Table 3, nor any asymmetry in the histogram in Fig. 1c can be observed for this score.

The $\tilde{q}^*_m$ values in Table 3 for different steps of the analysis show that the ability of the quality system to prevent human errors at the ICP–MS calibration ($\tilde{q}^*_{m=2}$) is minimal. The $\tilde{q}^*_{m=2}$ score values are

from 14% to 18% with a standard deviation varying from 4% to 8%, depending on the expert confidence. The variability of $\tilde{q}^*_m$ scores is the largest in comparison to other scores in Table 3. A difference of 1–2% between the mean and the median of the $\tilde{q}^*_m$ score values and an evident asymmetry of the histograms are observed. In particular, that is shown in Fig. 1d for the $\tilde{q}^*_{m=2}$ score.

Effectiveness of the whole quality system for all steps of the analysis is *Eff**=55%. It is a relatively low effectiveness in comparison to 59% for pH measurement of groundwater [4] and 71% for pesticide residue analysis [5]. The mean of the simulated *Eff** score values for the elemental analysis by ICP–MS is from 54% to 51% with a standard deviation varying from 3% to 5%. In general *Eff** tends to be overestimated in direct calculation from the elicited data, similar to *L** (opposed to *P** tending to be underestimated) when confidence of expert judgments is decreasing.

There is no difference between the *Eff** mean and median, and the score histogram in Fig. 1e is completely symmetric.

6.3. Robustness

The score robustness for the quality risk management and improvement of a laboratory quality system could be considered as satisfactory when a score relative variability, expressed as relative standard deviation RSD=STD/Mean, does not exceed 0.4 (rounded up from 1/3). In other words, a score is robust when variability (STD) of the expert judgments can be defined as insignificant in comparison to the score mean. Such a rule of 1/3 is used in metrology [31], e.g., for verification of weights [32] and preparation of test items for proficiency testing [33]. Practically the same rule is applied in spectroscopy for determination of limit of detection, as an analyte concentration equal to 3 STD of the measuring system response for a blank (noise) [34].

For example, for *Eff** score this requirement to robustness implies RSD=STD/*Eff** $\leq$ 0.4. In the case of the elemental analysis by ICP–MS the RSD of *Eff** is less than 0.1 for all models of the expert behavior. Therefore, one can assume that the robustness of this score is satisfactory. Other scores in Table 3 could also be assessed as robust enough. The $\tilde{q}^*_m$ scores, especially $\tilde{q}^*_{m=2}$, are less robust. However, even for an irresolute expert, the robustness of $\tilde{q}^*_{m=2}$ is still acceptable, since corresponding relative standard deviation is RSD=8/18=0.4.

It is important also that the score relative range (the difference between the maximal and the minimal score values related to their average), calculated from elicited data and simulated values does not exceed 0.4. From Table 3, the largest relative range is of score $\tilde{q}^*_{m=2}$, however this range is still acceptable, since (18–14)/16=0.25 < 0.4. Thus, the results of the human error quantification obtained in the case study are not dependent significantly on the kinds of calculation and simulation performed, i.e., are robust from this perspective as well.

7. Conclusion

The Monte Carlo method was applied for simulation of expert judgments on human errors in a chemical analysis and determination of score distributions for quantification of the errors. The simulation, based on modeling of an expert behavior by means of plausible pmfs, allowed evaluation of the score variability caused by variability of the expert judgments.

A case study of elemental analysis of geological samples by ICP–MS showed that in spite of achievements in instrument development there are still a number of human error scenarios, which should be taken into account in a routine laboratory for quality risk management.

The results of the human error quantification obtained in the case study are not dependent significantly on the kinds of calculation or simulation performed and can be assessed as robust for quality risk management and improvement of quality system in an analytical chemical laboratory.

Acknowledgements

This research was supported in part by the International Union of Pure and Applied Chemistry (IUPAC Project 2012-021-1-500). The authors would like to thank Prof. Yury Karpov (State Research and Design Institute for Rare Metal Industry, Russia), the project team member, for useful discussions.

References

[1] International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. Harmonized Tripartite Guideline ICH Q9: Quality Risk Management, 2005.
[2] ISO/TS 22367, in: Medical Laboratories—Reduction of Error through Risk Management and Continual Improvement, 2008. Technical Corrigendum 1, 2009.
[3] I. Kuselman, F. Pennecchi, A. Fajgelj, Y. Karpov, Accred. Qual. Assur 18 (2013) 3–9.
[4] I. Kuselman, E. Kardash, E. Bashkansky, F. Pennecchi, S.L.R. Ellison, K. Ginsbury, M. Epstein, A. Fajgelj, Y. Karpov, Accred. Qual. Assur 18 (2013) 459–467.
[5] I. Kuselman, P. Goldshlag, F. Pennecchi, Accred. Qual. Assur. 19, 2014, http://dx.doi.org/10.1007/s00769-014-1071-6.
[6] A.H. Perera, C.A. Drew, C.J. Johnson (Eds.), Expert Knowledge and its Application in Landscape Ecology, Springer, New York, 2012.
[7] P. de Barro, G. Gordh (Eds.), Plant Biosecurity Handbook, Springer, New York, 2011.
[8] V.M. Bier, Calibration of Expert Judgments in Counterterrorism Risk Assessment, Current Research Project Synopses, Paper 67, ⟨http://research.create.usc.edu/current_synopses/67⟩ (assessed in 2014).
[9] S. Dror, E. Bashkansky, R. Ravid, Int. J. Saf. Security Eng. 2/4 (2012) 317–329.
[10] L.H.G. Goossens, R.M. Cooke, Twenty years of experience with expert judgments, in: E. Zio, V. Ho Tsu-Mu Kao (Eds.), Int. Conf. on Probabilistic Safety Assessment and Management, Edge Pub Group Ltd, Hong Kong, 2008, pp. 1–8.
[11] R.M. Cooke, Expert Judgment Bibliography, Dept. Mathematics, T. U. Delf, ⟨http://bgconv.com/docs/index-83863.html⟩ (assessed in 2014).
[12] S. L. R. Ellison, A. Williams (Eds.), Eurachem/CITAC Guide: Quantifying Uncertainty in Analytical Measurement, third ed., 2012, pp. 24–25.
[13] R.M. Cooke, L.H.G. Goossens, J. Risk Res. 7 (2004) 643–657.
[14] J.C. Helton, M. Pilch, C.J. Sallaberry, Reliab. Eng. Syst. Saf. 124 (2014) 171–200.
[15] T. Kelly, Disagreement and the Burdens of Judgment, Princeton University, (assessed in 2014).
[16] The R Project for Statistical Computing, ⟨http://www.r-project.org⟩ (assessed in 2014).
[17] J. Maindonald, W.J. Braun, Data Analysis and Graphics Using R: An Example-Based Approach (Cambridge Series in Statistical and Probabilistic Mathematics), third ed., Cambridge University Press (2010) 82–84.
[18] JCGM 101, Evaluation of Measurement Data – Suppl. 1 to the "Guide to the Expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method, 2008, ⟨http://www.bipm.org/utils/common/documents/jcgm/JCGM_101_2008_E.pdf⟩.
[19] R.W. Murray, D.J. Miller, K.A. Kryc, Analysis of Major and Trace Elements in Rocks, Sediments, and Interstitial Waters by Inductively Coupled Plasma-atomic Emission Spectrometry, ODP Technical Note 29, 2000, ⟨http://www.odp.tamu.edu/publications/tnotes/tn29/index.htm⟩.
[20] US Geological Survey, Crustal Geophysics and Geochemistry Science Center, Solution ICP–MS Laboratory, ⟨http://crustal.usgs.gov/laboratories/icpms/solution.html⟩ (assessed in 2014).
[21] S.M. Eggins, J.D. Woodhead, L.P.J. Kinsley, G.E. Mortimer, P. Sylvester, M.T. McCulloch, J.M. Hergt, M.R. Handler, Chem. Geol. 134 (1997) 311–326.
[22] P. Dulski, Geostand. Newslett. 25 (2001) 87–125.
[23] COMAR, International Database for Certified Reference Materials, ⟨http://www.comar.bam.de/en⟩ (assesses in 2014).
[24] JCGM 200, International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM), 2008.
[25] Y.S. Liu, Z.C. Hu, M. Li, S. Gao, Chin. Sci. Bull. 58 (2013) 3863–3878.
[26] ISO Guide 33, Reference Materials—Good Practice in Using Reference Materials, third ed., 2014.
[27] Shi-WoeiLin, V.M. Bier, Reliab. Eng. Syst. Saf. 93 (2008) 711–721.
[28] M.S. Epstein, Talanta 80 (2010) 1467–1469.
[29] C.K.W. de Dreu, A. Evers, B. Beersma, E.S. Kluwer, A. Nauta, J. Organ. Behav 22 (2001) 645–668.

[30] S.P. Carmien, F.I. Cavallaro, R.A. Koene, 'Senior Moments': loss and context, in: PETRA '09 Proceedings of the Second International Conference on Pervasive Technologies Related to Assistive Environments, Article 44. ACM, New York, USA, 2009. http://dx.doi.org/10.1145/1579114.1579158.
[31] S.G. Rabinovoch, Measurement Errors and Uncertainties—Theory and Practice, third ed., Springer, New York (2005) 266.
[32] OIML R 111-1, International Recommendation, Weights of Classes E1, E2, F1, F2, M1, M1–2, M2, M2–3 and M3, Part 1: Metrological and Technical requirements, 2004, p, 11.
[33] I. Kuselman, A. Fajgelj, Pure Appl. Chem. 82 (5) (2010) 1099–1135.
[34] V. Thomsen, D. Schatzlein, D. Mercuro, Spectroscopy 18 (12) (2003) 112–114.