Limitations of the application of the Horwitz equation

Thomas P.J. Linsinger, Ralf D. Josephs

We revisit the basic assumptions of the Horwitz equation using the example of mycotoxin assays. Prediction intervals from the Horwitz equation often span one order of magnitude. Including this variation in calculation of Horrat values would lead to a range of values exceeding 2, which is often used as a criterion to assess interlaboratory comparisons, so we question the suitability of Horrat value for this purpose. In addition, available analytical data show significant improvement in reliability over time, which casts serious doubts on the applicability of the Horwitz equation for current analytical methods.

We discuss the use of the Horwitz equation in the analytical laboratory, and conclude that it is not suitable for estimating uncertainties, as required by ISO 17025.

The Horwitz equation can be a valuable summary of historical data of analytical performance. However, it should not be used as a performance criterion due to:

- **shortcomings of the basic model;**
- **uncertainty in the values determined using it; and,**
- **its incompatibility with accepted methods for the determination of measurement uncertainty, as required by ISO 17025.**

We recommend that, instead of using the Horwitz equation, there should be a proper identification of all components of uncertainty of measurement and reasonable estimation, as stipulated by the ''Guide to the Expression of Uncertainty in Measurements'' (GUM).

Keywords: Horrat value; Horwitz equation; Interlaboratory comparison; ISO 17025; Measurement uncertainty; Quality assurance

Thomas P.J. Linsinger*
EC-JRC, Institute for Reference Materials and Measurements (IRMM), Retieseweg 111, B-2440 Geel, Belgium

Ralf D. Josephs
Bureau International des Poids et Mesures (BIPM), Pavillon de Breteuil, F-92312 Sèvres Cedex, France

*Tel.: +32 14 571 956; Fax: +32 14 571 548; E-mail: thomas.linsinger@ec.europa.eu

1. Introduction

In 1980, Horwitz et al. published an evaluation of 1000 interlaboratory comparisons that led them to conclude that there is a fixed relationship between analyte level and reproducibility standard deviation (RSD_R) [1,2]. According to this analysis, the relationship between RSD_R and the analyte level c is:

$$RSD_R[\%] = 2^{(1-0.5\log_{10}c)} \quad (1)$$

irrespective of the kind of analyte, matrix or method. Equation (1) has been widely used to assess the quality of interlaboratory comparisons using the Horrat value, which gives a comparison of the actual precision measured with the precision predicted by the Horwitz equation for a measuring method at that particular level of analyte and is calculated as:

$$\text{Horrat} = RSD_{R,\,\text{measured}}/RSD_{R,\,\text{Horwitz predicted}} \quad (2)$$

A Horrat value of 1 indicates satisfactory interlaboratory precision, whereas a value of 2 indicates unsatisfactory precision (i.e. one that is too variable for most analytical purposes or where the variation obtained is greater than that expected for the type of method employed according to Horwitz).

Furthermore, the values predicted from the Horwitz equation have been used in legislation to set acceptance limits for analytical methods (e.g. [3]). It has even been suggested recently to use the results from the Horwitz equation as an estimation of measurement uncertainties [4]. Subsequent analysis of more datasets by Horwitz himself and others [5–7] frequently showed significant deviations from the values predicted by the original Horwitz function.

In this article, we will discuss some basic assumptions of the Horwitz equation and will compare results from mycotoxin interlaboratory studies spread over more than 30 years. Furthermore, we will point to the error made by using regression parameters in the prediction of future events. Finally, we will show why use of the Horwitz equation as a performance criterion contradicts modern practice in quality management and is not compliant with guides to the estimation of measurement uncertainties, such as the Eurachem Guide ''Quantifying Uncertainty in Analytical Measurement'' [8] and the ''Guide to the Expression of Uncertainty in Measurements'' (GUM) [9].

Mycotoxin data have been used because the original RSD_R values are available in the literature and more subsequent data are available for this group of analytes. However, this does not limit the generality of our conclusions, as data for other analyte groups show similar scatter.

Table 1. Regression parameters of the mycotoxin data from [5]

Parameter	Value
Slope ± *s*	1.291 ± 0.087
Intercept ± *s*	4.66 ± 3.01
Coefficient of determination (r^2)	0.218
Standard error of the estimation (s_{y*x})	26.80

2. Mathematical limitations of the Horwitz equation

In his original paper and in some of the follow-up papers, Horwitz demonstrated that the concentration level is the most important variable in explaining the reproducibility of interlaboratory-comparison studies. However, he never stated that all variation can be explained by this parameter, which is a crucial prerequisite for using the equation as a performance criterion.

Database 1 from Horwitz's paper on mycotoxin assays [5] was transferred to Statistica 7.0 in order to demonstrate this difference. A regression line of RSD_R versus $2^{(1-0.5\log_{10}c)}$ was calculated (Fig. 1). The regression parameters were calculated and are shown in Table 1.

Statistical regression reveals that data are homoscedastic and confirms the former statement of Horwitz et al. [5] that the mycotoxin assays in the period 1968–1992 show a higher variability than predicted from the Horwitz equation. The slope of the regression line RSD_R vs. $2^{(1-0.5\log_{10}c)}$ is significantly different from 1 on a 95%-confidence level (two-sided t-test), while the intercept is not significantly different from zero. This difference is also visible in Fig. 1, where the line from the classical Horwitz equation is completely outside the 95%-confidence band of the regression of the mycotoxin assays. However, more revealing is the coefficient of determination of 0.22. This indicates that only 22% of the variance of the RSD_R can be explained by the Horwitz equation. Although concentration certainly influences the reproducibility, nearly 80% of the variance remains independent of $2^{(1-0.5\log_{10}c)}$ and is not explained by the Horwitz equation. This poor fit is also reflected in the very broad prediction bands in Fig. 1.

This does not matter as long as the evaluation stopped at the regression (i.e. if the conclusion had been that the parameter ''concentration'' significantly influences comparability of results). However, the Horwitz equation has been abused to predict RSD_R values and to set performance criteria. To illustrate the fallacy of setting regression equal to prediction, RSD_R values for three different concentrations were calculated based on the regression parameters shown in Table 1. Furthermore,

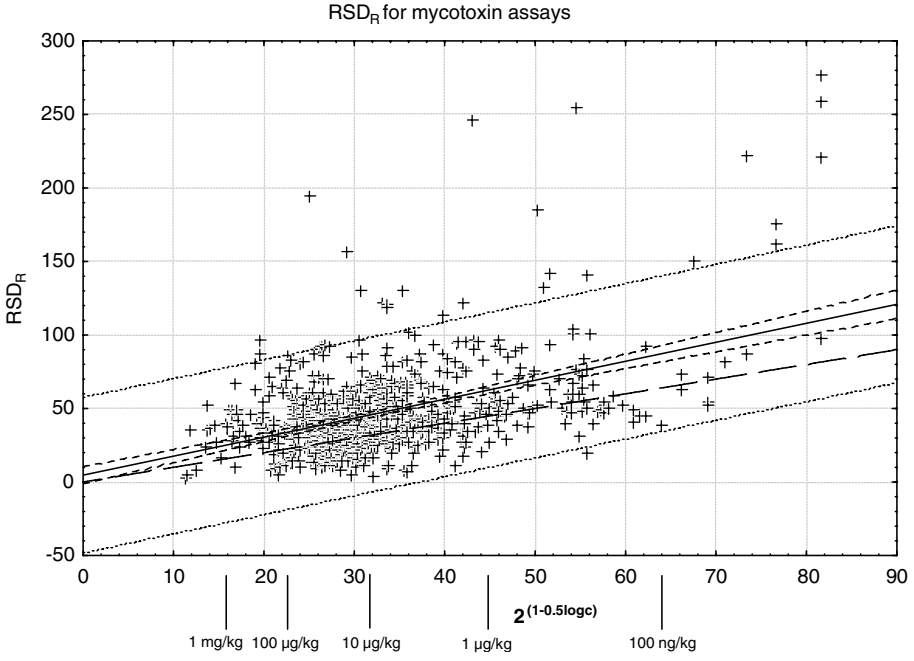

Figure 1. Mycotoxin data from [5]. Solid line: regression line for mycotoxins assays; inner dotted line: 95%-confidence interval of the regression line; outer dotted line: 95%-prediction interval for new observations; dashed line: regression line of the traditional Horwitz equation.

the half-width of the 95%-prediction intervals (PIs) for the respective RSD_R values were calculated according to Equation (3):

$$PI = t_{0.95,n-2} \cdot s_{y*x} \cdot \sqrt{\frac{n+1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \quad (3)$$

s_{y*x} standard error of the estimation
n number of data
x x-value for which the prediction interval is estimated
$\bar{x}$ average x+value of the regression line
x_i individual x-values of the regression line

These prediction limits are the limits between which an RSD_R for a given concentration can be expected. Predicted RSD_R values and their respective prediction intervals for three different concentrations are shown in Table 2.

By definition, PIs cannot be higher than their corresponding RSD_R values because this would result in negative RSD_R values. The occurrence of such cases (e.g., Table 2 for 10 μg/kg) clearly demonstrates the mathematical limitations of the Horwitz approach.

In general, severe deviations from the predicted RSD_R value can occur, as shown in Table 2. For a concentration of 15 μg/kg, more than 1 out of 20 intercomparisons will have an RSD_R of more than twice the predicted value. Transformed into the Horrat value, 1 out of 20 intercomparisons will have a Horrat value >2, which is generally regarded as unsatisfactory. The situation is even worse for higher concentrations because more than 12% of intercomparisons can be expected to have Horrat values >2 at concentrations above 200 μg/kg.

The situation is apparently better at lower concentrations (e.g. 0.1 μg/kg), where the upper 95%-prediction limit is ''only'' a factor of 1.6 above the predicted value. However, 0.1 μg/kg corresponds to an x-value of 64 in Fig. 1 and deviations from the predicted curve start to become particularly severe at this concentration level. Only at concentrations below 5.4 μg/kg is the 95%-prediction interval smaller than twice the RSD_R predicted. This means that only below this concentration is it statistically valid to state that Horrat values >2 are unsatisfactory. Horrat values >2 can be expected for all higher concentrations. However, the usefulness of predicting an RSD_R of 87 ± 52% (0.1 μg/kg) can be doubted.

Table 2. Calculated RSD_R and their confidence limits for mycotoxins

Concentration	$RSD_{R,\,predicted}$ ± PI (%)
$1 \cdot 10^{-10}$ (0.1 μg/kg)	87 ± 53
$1 \cdot 10^{-9}$ (1 μg/kg)	63 ± 52
$1 \cdot 10^{-8}$ (10 μg/kg)	46 ± 52

While the prediction error makes the use of Horrat values to judge the quality of intercomparisons doubtful, it invalidates its use for setting performance criteria, as the predictive uncertainty should be taken into consideration. Two examples demonstrate this:

- rather than stating ''at a mass fraction of 10 μg/kg, RSD_R should be 46% or better'', the correct statement should be ''RSD_R should be 46 ± 53%''; and, similarly,
- for a concentration of 200 μg/kg, the correct statement should be ''RSD_R should be 31 ± 52%''.

The limitations of this approach are obvious from the prediction bands, as shown in these examples. It furthermore suggests that ''anything goes'', a view that is unlikely to be taken by any legislator. If this is why the prediction intervals have not been used so far, one should not forget that not mentioning the uncertainty of the prediction does not eliminate it – it is just invisible. The Horwitz equation therefore does not allow method performance to be predicted with any reasonable certainty.

3. Methodological limitations of the Horwitz equation

In addition to the mathematical limitations, the Horwitz approach disregards other crucial prerequisites. Analytes, matrices, methods and time are considered as irrelevant for method reproducibility [10]. The approach is derived from regression analysis of many interlaboratory comparison studies, but is counter-intuitive and is contradicted by everyday experience in the laboratory.

3.1. Influence of analytes

It is a widespread observation that the kind of analyte affects repeatability and reproducibility of a measurement procedure. This is even the case for closely related analytes, such as trace metals. Evaluation of 13 proficiency tests of trace metals in water showed that the reproducibility of As is worse than that of Cd, even if the concentration levels of Cd are on average one order of magnitude below those of As [11]. Horwitz himself discussed the influence of the analyte [12], but this finding was afterwards neglected. Interestingly enough, Horwitz concluded in his paper on mycotoxins [5] that mycotoxin assays are less reproducible than other assays.

Apart from the intrinsic problems connected with some analytes, there can be another influence of the analyte; namely, that the concentration levels differ from analyte to analyte. Proficiency tests focus on analytes in typical concentrations. For the mycotoxins used in [5], this means that concentration levels for aflatoxins were

significantly below those of other mycotoxins, such as deoxynivalenol or zearalenone. Worse precision in the determination of aflatoxins can therefore be an intrinsic problem of aflatoxin determination rather than a concentration effect. This potential correlation between analyte type and analyte concentration casts further doubt on the applicability of the model.

3.2. Influence of methods

As was explicitly pointed out in [5], there is a difference in reproducibility between the various methods. Not surprisingly, measurements based on liquid chromatography (LC) tended to have a better reproducibility than those based on thin-layer chromatography (TLC), which in turn were more reproducible than enzyme-linked immunosorbent assay (ELISA) measurements. This shows that, as expected, the method clearly influences the precision of results, which is in strong contradiction with the basic assumption of the model.

In general, ELISA and TLC methods are used as screening tests for mycotoxins because of poorer performance in comparison with LC or GC methods [13]. A notorious problem of ELISA-based methods is cross reactivity of analog compounds resulting in large scatter and significant overestimation. For example, it is well-known that ELISA methods used for the determination of the mycotoxin deoxynivalenol (DON) principally cannot distinguish between the naturally-occurring mycotoxins DON, 3-acetyl-DON, 15-acetyl-DON, and 3,15-diacetyl-DON, because the mycotoxin antibodies are designed against 3,7,15-triacetyl-DON. These findings were underpinned by interlaboratory studies showing significantly higher results for ELISA methods, when compared with LC or GC results [14].

3.3. Influence of time

The passage of time should not in itself change the accuracy of measurements but analytical equipment changes with time. The combined effect of all these influences can be summarized in the variable ''time''. This influence will typically be more pronounced in new fields of analysis, be it new analytes or new concentrations. Improvement in trace analysis can therefore be expected to be larger than for long-established bulk analytical techniques.

The mycotoxin assays based on LC, TLC and ELISA evaluated in [5] span the period 1968–1992, i.e. nearly 25 years. It would be disappointing if the performance of laboratories had not improved over this time. Time can influence reproducibility of assays in many ways. One influence is certainly the availability of reliable reference materials, as shown in [15]. Lack and quality of calibrator and matrix reference materials was certainly a problem in the early days of mycotoxin analysis. Agreement between various laboratories might therefore have been expected to improve over time.

A second factor is the development of analytical equipment and methodologies. Looking at the database from 1993 [5], a vast number of interlaboratory studies was based on TLC and ELISA, which are today considered as semi-quantitative assays and screening tests at best. It might be expected that the wider use of LC and GC, the development of columns of better quality and more reliable standards would result in an improvement of the reproducibility over time, as also shown by Whitaker et al. [16]. This assumption is confirmed by statistical data. While in 1993, Horwitz et al. [5] concluded that mycotoxin assays tended to show reproducibilities above those predicted by the Horwitz equation, Thompson et al. showed that interlaboratory comparisons after 1997 have significantly better reproducibility than expected from the Horwitz equation, thus refuting an earlier statement that reproducibility does not improve over time [17].

A third factor is the improvement in analytical equipment itself. This improvement results frequently in higher signals that can be determined more accurately. Examples for this are modern diode-array detectors that are significantly more sensitive than older models. For example, this increased sensitivity of instrumentation can be exploited by narrowing the wavelength range that is measured, thus decreasing the influence of interferences. A major improvement in mycotoxin analysis was achieved by the technical advances in the field of LC-mass spectrometry (MS). In recent years very accurate, sensitive and robust LC-tandem mass spectrometry (MS^2) methods with electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI) have been developed for a variety of mycotoxins. These methods do not require any derivatization of the analytes, as GC methods and matrix effects are considerably reduced using the multiple reaction monitoring (MRM) mode [18].

In addition, the development of immuno-affinity columns or MycoSep columns for several mycotoxins improved sample clean-up [13,19]; for example, a recent interlaboratory study on aflatoxin B_1, B_2, G_1, G_2 mycotoxins in hazelnut paste by immuno-affinity column clean-up with LC – fluorescence detection (FLD) using post-column bromination demonstrated substantially improved RSD_Rs in the range 6.1–7.0% for total aflatoxins and 7.3–7.8% for aflatoxin B_1 at mass fraction levels of 4.0–11.8 μg/kg total aflatoxins [20] (predicted about 45%). Another example of significant improvement in method performance in the field of mycotoxin analysis was demonstrated in a project for the production of certified reference materials (CRMs) for aflatoxin M_1 (AfM_1) in milk powders [21]. The characterization study on AfM_1 in milk powders by different immuno-affinity column clean-ups with LC-FLD methods resulted in even better RSD_Rs in the range 5.9–4.8% [21] for AfM_1 at mass fraction levels of 0.11–0.44 μg/kg (predicted >70%), respectively.

4. Measurement uncertainty and the Horwitz equation

As described above, the predicted RSD_R value from the Horwitz equation has been prescribed as a criterion for method performance [3]. This means using it as an estimate for measurement uncertainty, as explicitly suggested by Massart et al. [4]. However, this practice is not in line with the GUM [9] and infringes more than one current quality-management practice.

Especially, the contradiction with the estimation of uncertainties of measurements as demanded in section 5.4.6.2 of ISO 17025 [22] is of utmost importance. The basic idea of any uncertainty estimation is to gather information on a particular measurement. In an ideal case, all uncertainty contributions for the particular measurement are evaluated. If measurement uncertainty is evaluated from method-validation data (e.g., as suggested by the Eurachem Guide [8]), the actual measurement must be linked to the average method performance using the normal quality-assurance tools such as quality-control charts. Using the Horwitz equation as an estimate for RSD_R means that there is no link whatsoever between the actual measurement and the uncertainty estimation. This means that, even if one accepts RSD_R values as estimates for uncertainties, the values from the Horwitz equation are most likely not to be accurate estimates for the measurements in question.

The ideal solution would be identification of all components of uncertainty of measurement and reasonable estimation, as stipulated by the aforementioned standards and guides [8,9,22]. In the field of mycotoxin analysis, considerable improvement in method performances was demonstrated in the project for the production of CRMs for AfM_1 in milk powders [21], as already mentioned above. Within the frame of this project, expanded measurement uncertainties for the determination of AfM_1 were also assessed for each laboratory by summing the combined standard uncertainties of sample mass, common calibrant, external calibration, precision and recovery at AfM_1 mass-fraction levels of 0.1 μg/kg and 0.4 μg/kg. Relative expanded uncertainties in the range 7.7–23.7% ($n = 7$) and 6.8–19.7% ($n = 8$) for the lower and higher mass-fraction levels were calculated. The relative expanded uncertainties are significantly better than anticipated from the Horwitz equation and, above all, the approach complies with the requirements of the GUM [9].

It might be argued that using the values from the Horwitz equation constitutes an ''expert judgment'', as explicitly foreseen in the GUM. However, the GUM clearly states that uncertainty evaluation requires ''detailed knowledge of the measurand and the measurement'' (3.4.8). Using data from other methods for other matrices and analytes is not in line with this requirement. Furthermore, it is difficult to argue that, if expected RSD_R values are somewhere between 11% and 115%, assuming a value of 63% is equivalent to an expert judgment.

Another incongruity in using the Horwitz equation as an estimate for measurement uncertainty concerns the demand of ISO 17025 that a laboratory must make efforts to determine the goal of the analysis of its customer. The result of a measurement always includes the measurement uncertainty, be it explicit or implicit. It is therefore assumed that the customer knows how accurate the results need to be. Using an average performance, such as the RSD_R from the Horwitz equation, corresponds to what is usual, but not to what is necessary and therefore complies with neither ISO 17025 nor other current approaches for judging laboratory performance. Thompson et al. [23] decidedly dismissed the practice of normalizing z-scores to the average performance in the recent IUPAC protocol for proficiency testing, as it does not take into consideration customer needs. This affects even more RSD_R values from the Horwitz equation, which correspond to outdated average laboratory performance, as shown above, and which are unrelated to current customer needs so they are no longer valid.

5. Conclusion

We have demonstrated that the fit of the individual data from the Horwitz equation is rather poor. The correlation is not good enough to use the Horwitz equation as a predictive model and confidence levels can exceed the expected values by more than a factor of 2.

In addition, there are technical reservations about the Horwitz equation (i.e. its assumptions that RSD_R values do not depend on analyte, matrix, method and/or time are mistaken, as demonstrated by recent interlaboratory studies). The Horwitz equation does not make allowances for the improvement of analytical methods and techniques, so using the Horwitz equation is benchmarking against outdated standards and leads to complacency with results that are not currently state-of-the-art.

Nevertheless, the Horwitz equation is a useful tool to summarize historical measurement performance but the unreliability of its results for a specific problem in question, the disagreement with modern quality management and the need of uncertainty estimation make it unsuitable as a criterion for method performance. Instead of using the Horwitz equation, measurement uncertainties should be identified and estimated according to the GUM approach [9], which is consistent with ISO 17025 [22].

In view of the movement away from specifying particular analytical methods towards specifying method performance criteria, the Codex Committee on Methods of Analysis and Sampling of the Codex Alimentarius Commission is discussing a fitness-for-purpose approach to evaluating methods of analysis [24,25]. This approach would be based on an uncertainty function constructed from precision data *inter alia*, and is to be judged against the Horwitz equation. This kind of uncertainty function should be reconsidered, because the Horwitz equation does not provide an accurate, traceable and state-of-the-art judgment, as we have shown above.

Acknowledgement

The authors were very saddened to hear about the death of William Horwitz while the present paper was being submitted. An eminent chemist and administrator at the Food and Drug Administration, he was the recipient of many prestigious awards for his work in analytical chemistry and was for many years Executive Director of the Association of Official Analytical Chemists (now AOAC International). We wish to acknowledge his important contribution to analytical food chemistry and his excellent work in the field of food standards.
We also thank Susanna Linsinger for transferring the data from Horwitz's publication into a spreadsheet.

References

[1] W. Horwitz, L.R. Kamps, K.W. Boyer, J. Assoc. Off. Anal. Chem. 63 (1980) 1344.
[2] W. Horwitz, Anal. Chem. 54 (1982) 67A.
[3] European Commission, Commission Decision 2002/657/EC implementing Council Directive 96/23/EC, Off. J. Eur. Commun. L221 (2002) 8.
[4] D.L. Massart, J. Smeyers-Verbeke, Y. Van der Heyden, LC-GC Eur. 10 (2005) 528.
[5] W. Horwitz, R. Albert, S. Nesheim, J. AOAC Int. 76 (1993) 461.
[6] W. Horwitz, R. Albert, J. AOAC Int. 79 (1996) 589.
[7] M. Thompson, Analyst (Cambridge, UK) 125 (2000) 385.
[8] S.L.R. Ellison, M. Rosslein, A. Williams, (Editors), EURACHEM/CITAC Guide Quantifying Uncertainty in Analytical Measurement, 2nd Edition, Eurachem, 2000 (http://www.eurachem.ul.pt/guides/QUAM2000-1.pdf).
[9] International Organization for Standardization (ISO), ISO Guide to the Expression of Uncertainty in Measurements, ISO, Geneva, Switzerland, 1995.
[10] W. Horwitz, R. Albert, J. AOAC Int. 89 (1996) 1095.
[11] W. Kandler, Aufbau und Betrieb eines Kontrollprobensystems zur Qualitätssicherung in der Wasseranalytik, PhD Thesis, Vienna University of Technology, Austria, 1999.
[12] W. Horwitz, J. AOAC Int. 84 (2001) 919.
[13] R. Krska, S. Baumgartner, R.D. Josephs, Fresenius' J. Anal. Chem. 371 (2001) 285.
[14] R.D. Josephs, R. Schuhmacher, R. Krska, Food Addit. Contam. 18 (2004) 417.
[15] R. Schuhmacher, R. Krska, J. Weingaertner, M. Grasserbauer, Fresenius' J. Anal. Chem. 359 (1997) 510.
[16] T. Whitaker, W. Horwitz, R. Albert, S. Nesheim, J. AOAC Int. 79 (1996) 476.
[17] M. Thompson, J.L. Philip, J. AOAC Int. 80 (1997) 676.
[18] F. Berthiller, R. Schuhmacher, G. Buttinger, R. Krska, J. Chromatogr. A 1062 (2005) 209.
[19] R.D. Josephs, R. Krska, Fresenius' J. Anal. Chem. 369 (2001) 469.
[20] H.Z. Senyuva, J. Gilbert, J. AOAC Int. 88 (2005) 526.
[21] R.D. Josephs, R. Koeber, T.P.J. Linsinger, A. Bernreuther, F. Ulberth, H. Schimmel, Anal. Bioanal. Chem. 378 (2004) 1190.
[22] International Organization for Standardization (ISO), ISO 17025, General requirements for the competence of testing and calibration laboratories, ISO, Geneva, Switzerland, 2005.
[23] M. Thompson, S. Ellison, R. Wood, Pure Appl. Chem. 78 (2006) 145.
[24] ALINORM 05/28/23, Report of the 26th session of the Codex Committee on Methods of Analysis and Sampling 2005, 28th session of the Codex Alimentarius Commission, Joint FAO/WHO Food Standards Programme, Rome, Italy, 2005. (www.codexalimentarius.net/download/report/636/al28_23e.pdf).
[25] CX/MAS 05/26/4, Proposed draft recommendations on the fitness-for-purpose approach to evaluating methods of analysis, 26th session of the Codex Committee on Methods of Analysis and Sampling, Codex Alimentarius Commission, Joint FAO/WHO Food Standards Programme, Budapest, Hungary, 2005. (ftp://ftp.fao.org/codex/ccmas26/ma26_04e.pdf).