codex alimentarius commission

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

WORLD HEALTH ORGANIZATION

JOINT OFFICE: Viale delle Terme di Caracalla 00100 ROME Tel: 39 06 57051 www.codexalimentarius.net Email: codex@fao.org Facsimile: 39 06 5705 4593

Agenda Item 3b **CX/MAS 05/26/4**
March 2005

JOINT FAO/WHO FOOD STANDARDS PROGRAMME

CODEX COMMITTEE ON METHODS OF ANALYSIS AND SAMPLING
Twenty-sixth Session
Budapest, Hungary, 4-8 April 2005

PROPOSED DRAFT RECOMMENDATIONS ON THE FITNESS-FOR-PURPOSE APPROACH TO EVALUATING METHODS OF ANALYSIS

BACKGROUND

The Codex Committee on Methods of Analysis and Sampling is responsible, amongst other things, for developing specific methods of analysis and endorsing those which have been submitted by various Codex Committees. It has developed General Principles for methods of analysis which have been included in the Codex Procedural Manual. It is also recommending that a criteria approach be developed for methods of analysis included in Codex Standards and, in association with that, is developing Working Instructions on the Implementation of the Criteria Approach for Codex Committees (see paper CX/MAS 02/5).

It has discussed, at the Twenty-fourth Session of CCMAS, two possible approaches to evaluating acceptable methods of analysis.

The two possible approaches to evaluating acceptable methods of analysis are:

- To identify specific performance parameters and assign numeric values to these (the traditional approach)
- To identify a "fitness-for-purpose" approach, taking all values into account by defining a single parameter – a fitness function.

This latter approach was discussed at the Twenty-Fifth Session of CCMAS, where it was agreed that: the Delegation of the United Kingdom would redraft the document on "fitness for purpose" with the assistance of a drafting group for further consideration. It was also noted that the fitness for purpose approach took all values into account by defining a fitness function as a single parameter. The document also defined the related uncertainty function, explained how the estimated characteristic function could be constructed from precision, and presented some examples of the application of this new procedure.

A number of delegations had reservations about the approach, but the Delegation of the United Kingdom was asked to revise the document for further consideration at the next session.

This approach is revised and given in Appendix I. It is given in the form of a scientific publication .

Some comments on this draft were received from the Delegation of New Zealand, and these are given in Appendix II.

RECOMMENDATION

It is recommended that the fitness-for-purpose approach be discussed at the Twenty-sixth Session of CCMAS. The Committee should then decide whether to develop it further within the framework of the CCMAS or to keep a "watching brief" on the international activities currently on-going in this area.

APPENDIX I: USING UNCERTAINTY FUNCTIONS TO PREDICT AND SPECIFY THE PERFORMANCE OF ANALYTICAL METHODS

ABSTRACT

In both European legislation relating to the testing of food and the recommendations of the Codex Alimentarius Commission, there is a movement away from specifying particular analytical methods towards specifying performance criteria to which any methods used must adhere. This 'criteria approach' has hitherto been based on the features traditionally used to describe analytical performance. This paper proposes replacing the traditional features, namely accuracy, applicability, detection limit and limit of determination, linearity, precision, recovery, selectivity and sensitivity, with a single specification, the uncertainty function, which tells us how the uncertainty varies with concentration. The uncertainty function can be used in two ways, either as a 'fitness function', which describes the uncertainty that is fit for purpose, or as a 'characteristic function' that describes the performance of a defined method applied to a defined range of test materials. Analytical chemists reporting the outcome of method validations are encouraged to do so in future in terms of the uncertainty function. When no uncertainty function is available, existing traditional information can be used to define one that is suitable for 'off-the-shelf' method selection. Some illustrative examples of the use of these functions in methods selection are appended.

DEFINITIONS OF TERMS USED IN THIS PAPER

Uncertainty function: algebraic relationship describing how uncertainty of measurement varies with the concentration of the analyte in the context of a specific analytical procedure applied to a defined class of test material.

Fitness function: uncertainty function that specifies levels of uncertainty regarded as fit for purpose.

Characteristic function: uncertainty function that describes the performance of a defined analytical procedure.

INTRODUCTION

It has been for many years the practice for specific methods of analysis to be prescribed in a number of sectors of application, for example, the food and feed sectors in both the European Union (EU) and the Codex Alimentarius Commission. However, it has been recognised that there are a number of disadvantages to that approach. In particular:

1) the analyst is denied freedom of choice and thus may be required to use an inappropriate method in some situations;

2) the fixed procedure inhibits the use of automation, and;

3) it is administratively difficult to change a method found to be unsatisfactory or inferior to another currently available.

A number of organisations, most notably the Codex Alimentarius Commission (CAC), have accepted an alternative approach comprising a set of criteria to which methods should comply without specifically endorsing specific methods that should be adopted. A number of papers have been prepared within the Codex system which outline the advantages and disadvantages of the traditional and the proposed present procedures.

1) The CAC has agreed that this "criteria" approach gives greater flexibility than the present procedure adopted by CAC and, with non-defining methods ("rational" methods), eliminates the need to consider and endorse several equivalent procedures. The CAC has recognised that the endorsement of many equivalent rational methods for any specific determination does cause confusion.

2) in some areas of food analysis there are many methods of analysis available that meet CAC requirements as regards method characteristics but are not considered because of time constraints.

3) the adoption of a more generalised approach would ensure that all applicable methods are brought into the CAC system and would not act as a disincentive for the adoption of new developments.

4) while it may be necessary to continue to prescribe a single reference method for a dispute situation, the criteria approach would certainly be applied to the present alternative methods.

The same approach has now been adopted by the EU, particularly in respect of contaminants in food, where specific methods of analysis are not prescribed but performance criteria are. There are already examples of these specifications for the determination of aflatoxins, patulin, trace elements, 3-MCPD and tin, while other proposals, PAHs for example, are currently in draft. The introduction of the criteria approach implies that the procedure of defining and quantifying the criteria required needs to be developed in a systematic manner. This is often complex and alternative approaches are discussed in this paper.

POSSIBLE APPROACHES TO EVALUATING ACCEPTABLE METHODS OF ANALYSIS

Two possible approaches to identifying acceptable methods of analysis are:

1) "the traditional approach", which relies on the identification of specific aspects of performance and assigning numeric values to these. The traditional indicators of performance are (a) accuracy, (b) applicability (matrix, concentration range and with preference given to "general" methods), (c) detection/determination limits if appropriate for the determination being considered, (d) linearity of calibration, (e) precision (repeatability and reproducibility), (f) recovery, (g) selectivity (interference effects etc), and (h) sensitivity. These are discussed in the Harmonised Guidelines[1], and have all been defined within the CAC system but are currently being reviewed.

2) a "fitness-for-purpose" approach, that takes everything into account by defining a single concept based on uncertainty – a fitness function. By using this concept analysts may (a) select an off-the-shelf method that, on the basis of the information available, will probably deliver acceptable (fit-for-purpose) results; and (b) demonstrate that the chosen method is, in fact, capable of delivering fit-for-purpose analytical results in the users laboratory. This 'fitness–for-purpose approach' applied to method selection is the main theme of this paper. Demonstrating that selected methods do fulfil requirements is the business of method validation, which is fully discussed elsewhere[1].

UNCERTAINTY FUNCTIONS

In principle, the only information that is required for method selection is a simple 'uncertainty function' comprising:

(a) a statement defining the analyte and the range of matrix types to which the method is to be applied;
(b) an algebraic expression $u = f(c)$ describing the relationship between the uncertainty of measurement and the concentration of the analyte; and
(c) the domain (concentration range) over which the function in (b) is applicable.

Of course, if the concentration range of interest is quite narrow, as is often the case for the output of highly-controlled industrial production, the uncertainty can be regarded as invariant and item (b) is specified by a single number.

Uncertainty functions can describe equally well the actual performance of a specific method (what has been called the 'characteristic function'[2]) and the uncertainty that is fit for purpose for a specific field of application (the 'fitness function'). In this context, the selection of a suitable method comprises the comparison of its characteristic function with the fitness function over the range defined in (c) above, a procedure that could be carried out numerically or graphically (Fig 1).

FITNESS FUNCTIONS

According to this scheme, before selecting a method, the analyst has to quantify the uncertainty that defines fitness for purpose. That is the uncertainty that minimises a loss function that balances the cost of analysis against potential losses due to incorrect decisions. While a formal decision-theory approach to that idea is possible[3], an uncertainty function could simply be agreed between the laboratory and the customer on the basis of experience and professional judgement, or might be defined by an agency representing a whole application sector of chemical analysis. As a simple example, the fitness function,

Eq. 1 $$u_f = 0.05c, \quad 0.1 < c < 5\ \%\ \text{m/m}$$

specifies that the standard uncertainty u_f should be 5% of the concentration *c* over the concentration range 0.1-5.0 % m/m. Once this fitness function has been defined, it can be used to judge whether the characteristic functions of particular documented methods are suitable. Subject to validation, a method is suitable for the application if it offers to provide an uncertainty that is lower than or equal to the fitness function over its whole defined concentration range. (If its characteristic function provides a *considerably*

lower uncertainty than the fitness function, however, it may be that the proposed method is unnecessarily accurate and therefore unnecessarily expensive.)

CONSTRUCTING CHARACTERISTIC FUNCTIONS FROM TRADITIONAL INFORMATION

While it is straightforward to define complete fitness functions, off-the-shelf methods are as yet seldom described by adequate characteristic functions. Hopefully that situation will change, but in the mean time we have to make do with the fragmentary and sometimes incomplete information provided by validation under the traditional headings listed above. We have to try to integrate into a single coherent uncertainty function such information as is provided under these traditional headings, together with our own judgements covering the uncertainty contributions caused by factors where no numeric information is available. The method advocated here is to build up an estimated characteristic function starting with a skeleton obtained from precision information.

Skeleton characteristic function based on precision

Precision is a useful starting point in estimating the uncertainty function, because the standard deviation of reproducibility σ_R accounts for a large measure, often the greater part, of the total uncertainty in a measurement. σ_R is the principal parameter estimated by a collaborative trial (interlaboratory method performance study) and is therefore often immediately available. Moreover, it is available as a function of concentration, because the collaborative trial is normally carried out with at least five different materials containing a range of concentrations of the analyte. We could reasonably estimate σ_R values at intermediate concentrations by interpolation. When σ_R values are available, other types of precision-related uncertainty are not separately required, because they are subsumed into the reproducibility. The aspects of uncertainty that are not included in σ_R (that is, method bias and matrix variability) have to be estimated separately (as shown below) and combined with σ_R in the appropriate way.

The main limitation of σ_R is that it is unavailable when the method has not been subjected to a collaborative trial, and likely to be under-estimated by within-laboratory precision studies. This situation can be ameliorated by the use of one or more surrogate estimates of σ_R. Firstly, method validations will always include simple estimates of repeatability standard deviation and/or run-to-run standard deviation. Repeatability standard deviation σ_r, estimated by within-run replication, can be converted into an estimate of σ_R by making use of the well-founded empirical observation that the expected value of the ratio

Eq. 2 $$\sigma_r / \sigma_R \approx 0.5 .$$

If run-to-run standard deviation σ_{run} is available (as it may well be either explicitly or implicitly in the form of internal quality control charts) that information can be used additionally or alternatively. While there is no comparably large body of experimental evidence to support it, an expected value of

Eq. 3 $$\sigma_{run} / \sigma_R \approx 0.8$$

is a reasonable presumption for current purposes.

Caution is required here because the naive methodology often used during method validation can give rise to low estimates of both σ_r and σ_{run}. For example, for a satisfactory estimate of σ_r, repeat measurements must be made on separate test portions of typical materials taken through the complete analytical procedure at random intervals throughout the typical duration of a routine run, preferably intercalated among normal test materials. Those precautions are often neglected during validation.

Any estimate of σ_R derived from lower-level precision experiments can be checked by reference to the Horwitz function,

Eq. 4 $$\sigma_H = 0.02 c^{0.8495} .$$

If the skeleton characteristic function is comparable with σ_H, the concurrence gives us confidence in the estimate. If the two functions differ systematically, expert judgement is required to choose between the estimates but, in the absence of sound evidence to the contrary, the higher indicated level of uncertainty is likely to be more correct.

Within-laboratory estimates of precision uncertainty are likely to be made at only one or two concentrations of the analyte. This information may have to be converted into a functional relationship over the range required. We can proceed by noting that that, at concentrations well above the detection limit, the relative standard deviation (RSD) of a method can often be regarded as a constant. So if only one such RSD is available, where the concentration is well above the detection limit, that RSD could be regarded as the initial characteristic function, at least over a limited concentration range. This would provide a characteristic function of the form

Eq 5 $$u_c = Ac,$$

where *A* is a constant. If two or more such estimates are available, an average of the RSDs could be used, again under the assumption that all analyte concentrations are well above the detection limit. Averaging would not be valid if the assumption of constant RSD were untenable, below say 50 times the detection limit.

Factors subsumed into the precision-based characteristic function.

Potential users of this idea may be worried that many traditional aspects of precision-related quality may be ignored in the foregoing set up. For example, tests for linearity, evaluation uncertainty derived from calibration data, systematic errors of calibration, and sensitivity are not mentioned. But their contributions to overall uncertainty are not ignored. Random calibration errors contribute to repeatability (within-run) variation and run-to-run precision. Systematic calibration errors (for example, those caused by incorrectly prepared stock solutions) are fully represented in reproducibility variation. Linearity is, of course, an important aspect of an analytical method, but bias due to lack of fit brought about by ignoring non-linearity would be represented in reproducibility variation. Sensitivity, the gradient of the calibration function, is for most analytical methods an essentially arbitrary quantity and plays no direct part in determining the uncertainty.

Taking account of the detection limit

Regardless of exactly how detection limit is conceptualised and defined, it represents the concentration levels where the net analytical signal is comparable with the magnitude of its uncertainty. Unless we are sure that we will be working well above the detection limit, we need to incorporate detection limit information into the characteristic function. That is easily accomplished. For example, taking the detection limit c_L as the concentration corresponding to a net signal of $\mu + 2\sigma$ produced by a test material containing no analyte, the standard uncertainty represented in the concentration domain would simply be numerically equal to $c_L/2$ (ignoring problems associated with the definition of uncertainty at near-zero concentrations). Combining this base-level uncertainty contribution with a putative proportional uncertainty present at higher concentrations (Eq. 5) provides a plausible and comprehensive model of uncertainty, with support from substantial amount of empirical data. By using the usual rule for combining independent uncertainties, this model gives us a skeleton characteristic function of

Eq. 6 $$u_c = \sqrt{c_L^2/4 + A^2c^2},$$

a form that has been noted experimentally in several studies[4].

When specifying the parameters of Eq 6, we must remember that detection limits quoted in the literature are usually 'instrumental detection limits', *i.e.*, they represent only instrumental precision under the best possible conditions and exclude any other, often much more substantial, sources of error. They are therefore unrealistically low and cannot be applied to real analytical systems without due consideration. For practical applications we need estimates of detection limits under reproducibility conditions. We should also note that Eq 6 contradicts the Horwitz function. This is because the Horwitz function is a generality about methods that are not near the detection limit.

(While detection limits may often affect characteristic functions, there is no general necessity for the fitness function to specify a baseline uncertainty at zero concentration. However, it will often be the case that there exists a level of uncertainty, below which it is unnecessary to go, however small the analyte concentration falls. In such instances, fitness functions of the form of Eq 6 can be employed.)

OTHER TRADITIONAL ASPECTS OF VALIDATION

The remaining traditional aspects of validation, that is, those not so far included in the skeleton characteristic function, are accuracy, applicability, recovery and selectivity. These factors are not independent, a circumstance that allows us to simplify the discussion. For example, accuracy depends on recovery and selectivity. Applicability comprises information about *inter alia* the types of matrix covered, which also bears on accuracy and has uncertainty implications.

Sometimes there is an uncertainty contribution caused by matrix variation *within* the defined scope of the method that has not been assessed or taken into account. An allowance for this deficit may be difficult to estimate, because the uncertainty contribution is seldom estimated in current validation practice. Therefore professional judgement may be called for in the estimation of the uncertainty contribution. Furthermore, if the proposed new use of the analytical method is *outside* the defined scope of the original validation, an additional uncertainty of unknown magnitude is introduced into the budget. Again, professional judgement would be required to estimate that contribution. This might be difficult. However, we must remember that, in the present context, these judgements are for method selection purposes only: the assumed uncertainties can be verified subsequently by validation experiments.

There is no general guidance as to whether these matrix effects should be regarded as translational or rotational, *i.e.*, additive or multiplicative. Again judgement is required for individual cases. If for example the matrix effect was judged to produce an extra multiplicative uncertainty of relative magnitude *B*, the adjusted characteristic function would take the form

Eq. 7 $$u_c = \sqrt{c_L^2/4 + \left(A^2 + B^2\right)c^2} \ .$$

Recovery information also has uncertainty implications[5]. Ideally, recovery factors (which are clearly measurements with uncertainties) should have associated uncertainty estimates. If these are available they should be combined into the characteristic function in the appropriate way. However, analysts should beware of double accounting here. Some recommended methods of estimating recovery factors might *include* contributions from matrix variation, for example if a variety of CRMs were used in the estimate. In addition, all uncertainty contributions separately estimated include the repeatability uncertainty.

REFERENCES

1. “Harmonised guidelines for single-laboratory validation of methods of analysis”, Michael Thompson, Stephen L R Ellison and Roger Wood, *Pure Appl. Chem.*, 2002, 74(5), 835-855

2. “Do we really need detection limits?”, M Thompson, *Analyst*, 1998, **123**, 405-407

3 “A decision theory approach to fitness for purpose in analytical measurement”, T Fearn, S Fisher, M Thompson, S R L Ellison, A*nalyst*, 2002, **127**, 818-824

4. “Variation of precision with concentration in an analytical system”, M Thompson, *Analyst*, 1988, **113**, 1579-1587

5. “Harmonised Guidelines For The Use Of Recovery Information In Analytical Measurement”, Michael Thompson, Steven L R Ellison, Ales Fajgelj, Paul Willetts and Roger Wood, *Pure Appl. Chem.*, 1999, 71, 337 – 348

ANNEX: EXAMPLES OF THE APPLICATION OF THE RECOMMENDED PROCEDURE.

Example 1. Short concentration range, well above detection limit, no collaborative trial data available.

Scenario
The analyte concentration is always in the range 40-50 % m/m.

The fitness function
The customer requires a standard uncertainty of 1.5 % m/m for fitness for purpose.

The relevant validation information available
The proposed analytical method provides the following, according to validation and IQC information. (All results are % m/m.)

- Repeat analyses of a typical test material (material 1) within run gives $\bar{x} = 41.6, s_r = 0.52$.
- The detection limit is estimated at 0.02.
- Use of two materials for IQC implied the following statistics:
  material 2: $\bar{x} = 25.3,\ s_{run} = 0.41$
  material 3: $\bar{x} = 52.3,\ s_{run} = 0.76$
- Analysis of ten spiked test materials estimated that the recovery of the analyte at the appropriate concentration was 95±2% relative.

Building the characteristic function

- As the relevant concentration range is small it is reasonable to regard the uncertainty as invariant.
- Material 1 is within the defined concentration range and by Eq. 2 gives us an estimate of $\hat{\sigma}_R = 0.52 / 0.5 = 1.04$.
- Material 2 is out of range and therefore should be ignored but, in fact, it provides an RSD similar to that of Material 3 (as expected when the concentration is well above the detection limit), and this adds confidence to the $\hat{\sigma}_R$ value derived from Material 3.
- Material 3 is just over range, so it might give an estimate somewhat on the high side, but in fact (by Eq. 3) gives $\hat{\sigma}_R = 0.76 / 0.8 = 0.95$.
- The Horwitz function (Eq. 4) at a concentration of 50% m/m gives a reproducibility estimate of $\sigma_H = 1.1$, which is consistent with the above $\hat{\sigma}_R$ estimates and reinforces our confidence in them.
- The detection limit is far below the required range so the zero-point uncertainty makes a negligible contribution to the uncertainty and is ignored.
- We therefore use a consensus of the concordant results for material 1 and material 3 to give $\hat{\sigma}_R = 1.0$ as the skeleton function.
- We build into the uncertainty an allowance for the uncertainty on the recovery factor. The uncertainty expected on the recovery-corrected mid-range result (45% m/m) is therefore $u_c = 45\sqrt{\left(\frac{1.0}{45}\right)^2 + \left(\frac{2}{95}\right)^2} = 1.4$.

Comparison of uncertainty functions
As $u_c < u_f$, the method is deemed suitable.

Example 2. Extended concentration range, no collaborative trial data.

Scenario
A commodity is being sold on the basis that the concentration of a particular constituent falls between 5 and 50 ppm. Experience has shown that batches analysed can have levels of the constituent over a somewhat wider range than that.

The fitness function
Preliminary considerations suggest that an RSU of about 10% over the range indicated, and down to 2 ppm, would meet requirements, so the fitness function is $u_f = 0.1c, \qquad 2 \le c \le 50$.

The relevant validation information available about the candidate method

- The instrumental detection limit of the method gleaned from the literature was 0.25 ppm when adjusted for the dilution of the test portion after chemical treatment.
- The repeatability standard deviation, estimated from 10 individual test portions of two putatively typical test materials in a single run, was reported as follows:
  $\bar{x}_A = 31, \qquad s_A = 1.4, \qquad RSD = 0.045;$
  $\bar{x}_B = 103, \qquad s_B = 3.4, \qquad RSD = 0.033.$
- Single determinations of the analyte in five certified reference materials provided the following results.

Certified value	Uncertainty on certified value	Reported value
8.0	0.4	7.2
108	5.1	101.3
21.5	0.8	22.7
42.1	0.9	40.9
20.2	1.0	20.4

Building the characteristic function

- Realistic detection limits are often much higher than instrumental values, and we adopt the arbitrary decision to raise it by a factor of five to 1.25 ppm. (This is consistent with what little information exists.) As a baseline standard deviation, its contribution is 1.25/2 = 0.625, and that may not be negligible at the low end of the concentration range of interest, so it will be included in the model.
- The RSDs of the two materials are comparable, and the lower concentration is apparently about 100 times the detection limit, so it is a reasonable assumption that both are estimates of the same constant RSD, so we adopt the mean of the two RSDs (namely 0.039) multiplied by 2.0 as the likely value of *A* in Eq. 5, 6 etc. (Note: the factor of 2.0 is derived from the application of Eq 2, because the reported standard deviations were estimated under repeatability conditions.) Under these assumptions, the skeleton characteristic function is given by

  $$u_c = \sqrt{0.625^2 + 0.078^2 c^2}\,.$$

- This is reasonably consistent with predictions from the Horwitz function. For example, at 10 ppm the predicted standard uncertainties are $u_c = 1.0$ and $\sigma_H = 1.1$, while at 50 ppm $u_c = 3.9$ and $\sigma_H = 4.4$ ppm. Accordingly we feel confident in using the skeleton function.
- Considering the results on the reference materials, a reasonable way of summarising them is obtained by looking at the relative differences between the certified values and the found values, i.e., $\left(x_{found} - x_{cert}\right)\big/x_{cert}$ which gives the following (in the tabulated order): -0.10; -0.06; 0.08; -0.03; 0.02. If the mean of these values were significantly different from zero (indicating evidence of overall bias) or the standard deviation were substantially greater than the relative uncertainty expected, the result would suggest that there was an additional source of uncertainty, perhaps due to matrix variation, that needed to be taken into account. In fact the mean relative deviation is zero and the standard deviation is 0.053. The expected relative standard uncertainty (RSU) is obtained by combining the mean RSU of the certified values (0.04) with the relative standard deviation of run-to-run precision,

which is estimated as 0.8 × 0.078 = 0.062. Thus the expected RSU is $\sqrt{0.053^2 + 0.062^2} = 0.08$. This is greater than the observed value, so there are no compelling grounds for inflating the skeleton function. (In fact we might harbour suspicions that the results on the reference materials were unnaturally accurate.)

- This gives us a final characteristic function of

$$u_c = \sqrt{0.625^2 + 0.078^2 c^2}\ .$$

Comparison of uncertainty functions

Fig. 1 shows the plots of u_c and u_f. Over most of the designated concentration range, the characteristic function is below the fitness function, showing that the proposed analytical method could apparently deliver the required degree of accuracy. However, at concentrations below about 10 ppm, the characteristic function is too high, as a consequence of the baseline variability. In order to meet the fitness-for-purpose requirement over the whole range, a method with a lower detection limit would be required.

Example 3. Extended concentration range, collaborative trial data available.

Scenario

The requirement addresses the determination of a trace constituent that usually occurs at concentrations between 10 and 100 ppm. From general experience with similar tasks, the nature of the test material and of the proposed method of determination together suggest that uncertainty due to matrix variation within the defined class would be restricted to less than 5% of the concentration. Recovery is expected to be 100%, because there is no scope for loss of analyte.

The fitness function

The client specifies that the uncertainty on the result should not exceed 10% of the concentration or 5 ppm, whichever is the greater.

The relevant validation information available

A collaborative trial has been carried out and the results are as follows.

Test material	Concentration	σ_R
A	16.0	1.2
B	31.4	2.0
C	39.8	2.5
D	42.9	3.7
E	46.6	3.8
F	57.1	3.4
G	63.2	4.4
H	69.9	4.0
I	88.6	7.1
J	94.3	5.1

Building the characteristic function

A plot of the collaborative trial data shows the trend of the σ_R as increasing with the concentration of the analyte (Fig.2). Fitting the data to Eq. 6 provides the estimates $\sigma_R(0) = 1.3$, $c_L = 2.6$ and $A = 0.064$. The resulting relationship, the skeleton characteristic function, is given by

$$u_c = \sqrt{1.68 + 0.0041c^2}\ .$$

(The Horwitz function is shown for comparison, and predicts somewhat higher σ_R than observed. In this instance we can ignore the discrepancy and simply use the collaborative trial data.) We now incorporate an extra rotational uncertainty to account for matrix variations equivalent to the maximum thought likely in this analytical system, that is, 5% relative, which, using Eq.7 gives the final characteristic function

$u_c = \sqrt{1.68 + \left(0.05^2 + 0.0041\right)c^2}$ or $u_c = \sqrt{1.68 + 0.0066c^2}$.

Comparison of uncertainty functions

The characteristic function and the fitness function are shown in Fig. 3. The characteristic function is seen to be the lower over the whole of the relevant range (10-100 ppm), so the method is apparently fit for purpose.

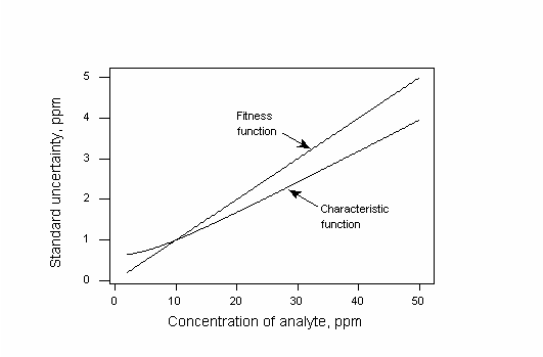


Fig 1. Fitness and characteristic functions, Example 2.

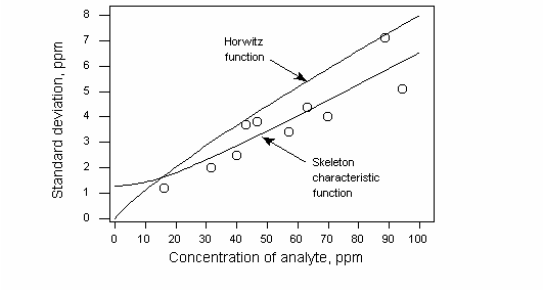


Fig.2. Collaborative trial results (circles) from Example 3, with fitted characteristic function and Horwitz function.

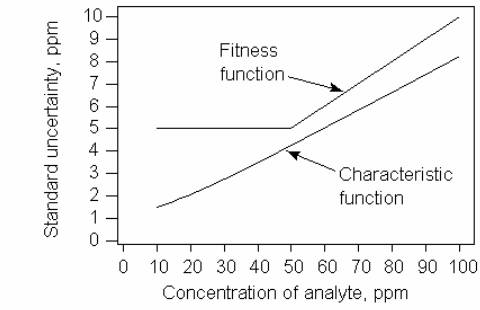


Fig.3. Fitness and characteristic functions, Example 3.

APPENDIX II: NEW ZEALAND COMMENTS ON THE FITNESS-FOR-PURPOSE APPROACH TO EVALUATING METHODS OF ANALYSIS

Summary

1. This paper offers comments on the "fitness function" as a means of assessing fitness for purpose, and suggests an approach by which the fitness of a method of analysis for the purpose of assessment of product conformance may be judged by estimating the additional product assessment error introduced by the measurement error of the method.

Introduction

2. Governments use methods of analysis of foods for a range of purposes, but most commonly to judge the conformance of a food product to a specified criterion.

3. Codex recommends that methods of analysis should either be selected from lists of methods that have been endorsed by CCMAS, or alternatively methods may be selected if they satisfy specified criteria for the performance of the methods.

4. Method performance is multidimensional, including repeatability, reproducibility, bias, sensitivity and so forth, and therefore requirements for endorsement or specification of required performance may include requirements on all, or several of, these parameters. The Draft Guidelines for Evaluating Acceptable Methods of Analysis[1] proposes means of judging the acceptability of methods of analysis using these parameters.

5. Other means of judging methods of analysis are being considered: the use of uncertainty functions offers a simplified approach; alternatively a candidate method may be considered in the context of its use in assessing product conformance.

Judging Fitness Using Uncertainty Functions

6. The paper "Using Uncertainty Functions to Predict and Specify the Performance of Analytical Methods" puts forward the opinion that detailed specifications of method performance are unnecessary, and that the specification of the "uncertainty" (an undefined term) at various concentrations of analyte is all that is required. This specification is called a fitness function.

7. The paper also puts forward the view that a laboratory may reasonably demonstrate (to its own satisfaction, which appears to be all that is needed) compliance of a method with such a criterion using a less extensive investigation than the collaborative trials traditionally required. The technique involves the use of various rules of thumb including the Horwitz function and "average" values for the ratios of the various error standard deviations. The paper does not put forward a convincing case that the rules of thumb give estimates as good as a collaborative trial.

8. The paper also appears to suggest that the traditional considerations involved in method validation are superseded by this approach: that, given the uncertainty function, detailed consideration of such things as bias, sensitivity and non-linearity is no longer necessary.

9. Several delegations expressed reservations about the "fitness function" approach at CCMAS25[2], and we offer a detailed commentary (see Annex 1). We consider that this approach is definitely not acceptable, and is suitable as an emergency procedure at best, in situations where only a reproducibility-like uncertainty is available for the method.

Judging Fitness for the Purpose of Assessing Product Conformance

10. Neither the traditional approach nor the "fitness function" approach demonstrates that a candidate method is actually fit for purpose. In the traditional approach there is no consideration of its purpose in the evaluation of foods. In the "fitness function" approach there is no consideration of purpose in the formulation of the uncertainty function that the fitness function is to constrain. It seems that there is potential for an increase in disputes, caused by the unsuspecting use of methods that are not really fit for purpose.

---

1 ALINORM 04/27/23, Appendix V.

2 ALINORM 04/27/23, paras 59 – 65.

11. In our view, fitness for purpose should be based on an overall evaluation of the measurement system in the context of the task in hand, which for Codex purposes usually involves the assessment of conformance of some parameter to some limit.

12. To evaluate fitness of a method, it is necessary to go beyond its performance characteristics and consider the effect of the measurement system on the decisions made using the results produced by the method. In the assessment of product conformance, these indicators relate to risk and possibly cost.

13. In product assessment applications there are two sources of uncertainty: sampling error and measurement error. In many cases the operating characteristics of decision rules have been calculated and approved as a function of the sampling protocol and the "state" of the lot to be assessed, possibly expressed as percent defective, but without taking account of measurement error. The practical consequence of the existence of measurement error is either to increase the risks from those calculated, or to incur increased costs either through increased sampling or because of changes to process parameters to accommodate new decision rules. Examples of the additional risk introduced by the measurement error of a method are shown in Annex 2.

14. We propose therefore that assessing fitness for purpose should involve the estimation of the additional risk or cost introduced by the measurement error of the method. The additional risk can be evaluated statistically.

Conclusions

15. It is recommended that CCMAS should not adopt the “fitness function” as a means for assessing the fitness for purpose of methods of analysis.

16. It is recommended that CCMAS consider the alternative approach of judging the fitness of a method for its purpose by estimating the additional product assessment error introduced by the measurement error of the method.

ANNEX A: COMMENTS ON THE PAPER, "USING UNCERTAINTY FUNCTIONS TO PREDICT AND SPECIFY THE PERFORMANCE OF ANALYTICAL METHODS"

Summary

The methods recommended seem appropriate only in very restricted circumstances, yet are put forward as universally applicable. No critical appraisal is given of circumstances in which they may or may not be appropriate, and even the concept of "uncertainty," fundamental to the paper, is not clearly, or even explicitly, defined. The paper could easily be read as supporting a reduction of scope and stringency in method validation, and as an encouragement to laboratories to use methods, at their own discretion, whose characteristics are quantified by data that would not normally be considered adequate.

The paper proposes that a single number, possibly varying with concentration, can summarise the performance of an analytical method. To expect that such a number, used alone, could be adequate to evaluate fitness for any purpose for which estimation may reasonably be required, is futile. Different purposes require different mixtures of the various components of error.

The sufficiency of the uncertainty function

With regard to the main point of the paper, that a reproducibility-like uncertainty function is adequate to judge fitness for purpose, the authors seem firstly to overstate their case, and secondly to provide no evidence in support of it.

On p3, "Uncertainty functions" the authors begin "In principle, the only information that is required for method selection is a simple uncertainty function ….." It would be interesting to know the principle to which they are referring.

An example where it could provide the only information required is the testing of a single sample of material (possibly a composite) against a cut-off. Here the appropriate cut-off could be calculated, and the probabilities of rejection at various concentrations calculated, using only the reproducibility (assuming the method bias and its uncertainty are both negligible.)

However there are certainly purposes for which the reproducibility-based uncertainty that the authors recommend is inappropriate:

- Even the next simplest case, where two samples are taken and non-compliance is found where one exceeds the cut off, requires repeatability and reproducibility net of repeatability to be known separately, just to calculate the appropriate cut-off.
- Even for such a simple purpose as estimating the mean concentration of a number of samples a simple estimate of reproducibility is not adequate to assess the precision of the resulting mean.
- The Codex requirement that butter should contain less than 16% moisture is monitored in New Zealand by a "variables sampling plan" (as documented by Codex). Evaluation of the characteristics of this sampling plan cannot be carried out if only reproducibility is available.

In fact reproducibility, or any other single measure of uncertainty, contains the various components of error in a certain mix. A single mix will not be suitable in all circumstances.

If only a reproducibility-like uncertainty is available for a method, we are reduced to using the rules of thumb listed by the authors for this type of calculation. This is suitable as an emergency procedure at best. The authors do not put forward a convincing case that the rules of thumb give estimates as good as a collaborative trial. There is apparently no logic behind the rules of thumb; they have just been found to work reasonably well in practice in most circumstances. To accept such rules as a satisfactory substitute for the use of estimates obtained in validation trials is not scientifically justifiable.

Simplification of method validation

The abstract (p 1) seems particularly unfortunate. It appears to exhort analysts to abandon the traditional quantification of method parameters in favour of what is essentially a less informative procedure, not for reasons of cost and convenience, but as a purported methodological advance.

The proposed fitness function is arbitrary so that any other, equally valid fitness function could be proposed, with the possibility that a method found fit by some functions might not be fit by others.

The construction of the fitness function allows a trade-off between different characteristics. For instance two alternative methods could be found acceptable but one could have a small bias and large imprecision and the

other a large bias and small imprecision. However the two candidate methods might not be "equal" in their impact on the compliance test involved. Moreover bias and precision are not tradable with the limit of detection – the former are performance characteristics and the latter more a "quality assurance" characteristic. It is also possible that the tolerances provided for some method characteristics in the fitness function might mean that other characteristics become irrelevant in the fitness assessment.

The acceptance level attached to the fitness function should take account of the estimation uncertainty of each of the characteristics included in the function, for a fair and valid decision to be made whether the candidate method 'conforms'. This highlights another problem with fitness functions: they provide no immediate diagnostic information about where a candidate method may not meet the requirements. Such information is desirable so that method improvement can be undertaken, or suitable allowances introduced.

The paper states (p6: Factors subsumed into the precision-based characteristics) that traditional sources of error are not ignored but get incorporated into the reproducibility. This is a very loose statement and needs clarification. The situation may be best explained using an example. The average measured value (over many labs) is supposed to be proportional to true concentration. The constant of proportionality is supposed to be unity. If it is not we get a method bias varying with concentration. This does **not** affect the estimated reproducibility. If there is some uncertainty about the constant of proportionality the bias will also be uncertain. The corresponding uncertainty is **not** included in reproducibility. If the constant varies from laboratory to laboratory, or from run to run within a laboratory, or from sample to sample within a run, this variation **does** get incorporated into the (relevant component of) reproducibility. Such variation would be a contributing cause to the uncertainty in the overall coefficient, and if there are a good number of laboratories in the trial, the uncertainty in the method bias will usually be swamped by this contribution to estimated reproducibility, and thus can be ignored. But the method bias itself will remain, unallowed for in the reproducibility. And if the number of laboratories in the trial is small, the uncertainty in the method bias will not be swamped.

It is seen that there is still a need to examine the constant of proportionality at the method level, and probably, unless the number of laboratories is large, the uncertainty attached to it as well.

Method bias

In the above it was a component of method bias that caused the problem. In fact the whole question of method bias is considered very cursorily if at all in the paper. The only explicit reference to it is (p4, last sentence,) "The aspects of uncertainty that are not included in sigma R (that is method bias and matrix variability) have to be estimated separately (as shown below) and combined with sigma R in an appropriate way." The method of combining method bias is not shown, and it is hard to see how it could be combined with sigma R in an appropriate way. One would think it would have to be applied as an additive adjustment to the estimates themselves. Presumably it is assumed to have been removed under a method calibration. However, this may not always happen, particularly in the absence of a gold standard, and if it happens, it may not happen successfully.

For example, the Direct Fat measurement of fat in butter seems at present to underestimate fat content by, on average, about 0.15 pp relative to the traditional methods. This is significant in at least one context in which the test is used. But which method is correct? No one knows. Presumably all methods should, by logic and chemistry, yield the correct result. There is no suggestion that the Direct Fat method should be modified to incorporate an ad hoc automatic adjustment of 0.15 pp, or that the other methods should be modified in the other direction. The suggestion would seem to be that a laboratory should be free to choose Direct Fat or one of the standard methods at will provided that that the methods all have similar reproducibility. This is unacceptable.

As noted above, if the number of laboratories involved in a validation study is small, and particularly where only a single laboratory is involved, as appears possible under the proposals in the paper, it is not only bias itself that causes problems but the uncertainty surrounding its estimation. The existence of a statistically significant method bias when a method is validated in a collaborative trial would be expected to call for investigation, at least unless the bias is small compared to the between-laboratory error net of run and repeatability effects. (**not**, one hopes, the reproducibility.) However the absence of a statistically significant bias does not prove that a bias does not exist. It merely sets limits (which may vary with concentration) on its likely size. The width of these limits gives rise to an uncertainty which may need to be taken into account.

It is certainly agreed that the reporting of these limits, or some "uncertainty" based on them, would be desirable in a validation study, but this does not seem to be included in the uncertainty that the authors mean a validation study to report instead of the "traditional information." Indeed, it appears not to have been considered, and it is not allowed for in their examples.

In the case of a substantial collaborative trial involving many laboratories, uncertainty caused by method bias will probably turn out to be negligible compared to the between-laboratory error. In the case of the ad hoc investigations of off-the-shelf methods, carried out by a single laboratory, that the authors appear to suggest, it almost certainly will not be negligible. This uncertainty, which is neglected by the authors, is of comparable magnitude to the effects taken into account.

Methods of estimating the uncertainty function

Some of this material could be quite useful in a different context. Methods of fitting plausible curves to uncertainties estimated at a few concentrations from a collaborative trial or elsewhere are discussed.

The following rules of thumb, stated to have good empirical foundation, are put forward.

1) Sigma R = 2 sigma r (i.e. reproducibility s.d. is twice the repeatability s.d.)
2) Sigma(run) (variation between duplicates in different runs within a lab) = 0.8 sigma R.
3) The Horwitz function

A point that needs to be made about these rules is that repeatability and within- laboratory reproducibility, as estimated in collaborative trials, will normally be average values. Some laboratories will do better and others not so well. Although the rules may work reasonably well in relating the averaged values to reproducibility, assuming that they will hold for any particular laboratory to the extent that it can predict method reproducibility from an estimate of **its own** repeatability may be rather a large assumption.

In fact, when used for estimation rather than checking, as Examples 1 and 2, they would have to be regarded as extreme measures.

For instance, in example 2 an estimate of reproducibility is made from two estimates of repeatability with 18 degrees of freedom in total, with all results taken from a single session. A possible method bias is discounted on the basis of data giving a 95% confidence interval for the relative bias of –14% to 14%. Apparently such a potential bias would not disqualify the method when a maximum relative uncertainty of 10% is required. Presumably the example is simplified in order to exemplify the techniques used. If so, this should certainly be made clear.

Codex should be very reluctant to publish material that authorizes, or appears to authorize, estimates of method reproducibility based on data from a single laboratory.

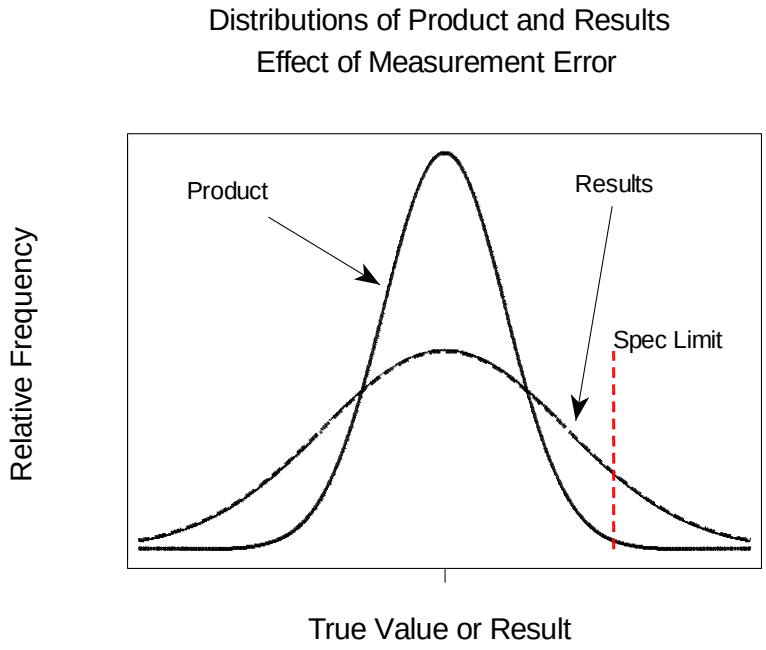
Further matters

- Two valuable points made in the paper are the need to consider uncertainty and its components as possibly related to concentration (throughout), and the requirements for valid estimates of repeatability (p5, paragraph following Equation 3.)

- The authors state that specification of a single reference method may be necessary in a dispute situation. However in a regulatory context, every case is potentially a dispute situation. A dispute situation will develop every time a non-conformity is alleged on the basis of the analytical results. Is it common for the initial result to be completely discarded, and the material under dispute to be re-analysed under a reference method, in these circumstances?

ANNEX B: ESTIMATION OF THE ADDITIONAL RISK INTRODUCED BY THE MEASUREMENT ERROR OF A METHOD

In general, unless specific allowance is made, imprecision of a test method will increase the risks of failing product of acceptable quality product and of accepting poor quality product. The first figure shows that the additional variability introduced by the test method causes the proportion of results above the upper specification limit to appear greater than for the product itself.



The second figure shows the converse. In a situation where the product is largely out of spec against the upper limit, measurement error will make it appear that a greater proportion of results will appear in spec than is actually the case.

Distributions of Product and Results
Effect of Measurement Error