Pure & Appl. Chem., Vol. 65, No. 9, pp. 2123–2144, 1993.
Printed in Great Britain.
© 1993 IUPAC

INTERNATIONAL UNION OF PURE
AND APPLIED CHEMISTRY

ANALYTICAL, APPLIED, CLINICAL, INORGANIC AND
PHYSICAL CHEMISTRY DIVISIONS
INTERDIVISIONAL WORKING PARTY FOR HARMONIZATION OF
QUALITY ASSURANCE SCHEMES FOR ANALYTICAL LABORATORIES*

THE INTERNATIONAL HARMONIZED PROTOCOL FOR THE PROFICIENCY TESTING OF (CHEMICAL) ANALYTICAL LABORATORIES

(Technical Report)

Resulting from the Symposium on Harmonization of Quality Assurance
Systems in Chemical Analysis, Geneva, Switzerland, May 1991
held under the sponsorship of IUPAC, ISO & AOAC

Prepared for publication by

MICHAEL THOMPSON[1] and ROGER WOOD[2]

[1]Department of Chemistry, Birkbeck College (University of London), London WC1H 0PP, UK
[2]Food Science Laboratory, Norwich Research Park, Colney, Norwich NR4 7UQ, UK

*Membership of the Working Party during 1989–92 was as follows:

Chairman: M. Parkany (Switzerland); *Members*: P. de Bièvre (Belgium); S. S. Brown (UK); L. E. Coles (deceased) (UK); R. Greenhalgh (Canada); B. Griepink (Belgium); A. Head (UK); R. F. M. Herber (Netherlands); W. Horwitz (USA); S. H. H. Olrichs (Netherlands); G. Svehla (Ireland); M. Thompson (UK); R. Wood (UK).

Republication of this report is permitted without the need for formal IUPAC permission on condition that an acknowledgement, with full reference together with IUPAC copyright symbol (© 1993 IUPAC), is printed. Publication of a translation into another language is subject to the additional condition of prior approval from the relevant IUPAC National Adhering Organization.

The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories (Technical Report)

Synopsis

The International Standardising Organisations AOAC International [then the Association of Official Analytical Chemists], ISO and IUPAC have cooperated to produce an agreed protocol on the "Design, Conduct and Interpretation of Collaborative Studies"(1). The Working Group which produced the protocol met and agreed at its April 1989 Washington D.C. meeting to develop a further protocol on proficiency testing, i.e. consideration of the use of results generated in interlaboratory test comparisons for the purpose of a continuing assessment of the technical competence of participating testing laboratories.

Such a harmonised protocol would have to outline the minimum requirements for agencies (laboratories or other organisations) who wished to develop and operate proficiency testing schemes and recommend statistical treatment of the reported data.

A draft harmonised protocol for the organisation of proficiency testing schemes was prepared and discussed at the AOAC International/ISO/IUPAC Meeting on the "Harmonisation of Quality Assurance Systems in Chemical Analysis", Geneva, May 1991, as part of the development process of such a protocol, and finalised at a meeting of the Working Party in Delft, The Netherlands, May 1992.

CONTENTS

1. INTRODUCTION

For a laboratory to produce consistently reliable data it must implement an appropriate programme of quality assurance procedures.

Analytical methods must be validated as fit for their purpose before use in the laboratory. Whenever possible validation should be achieved by means of collaborative trials that conform to a recognised protocol[1]. These methods must be fully documented, laboratory staff trained in their use and control charts should be established to ensure that the procedures are under statistical control. Where possible, all reported data should be traceable to reliable and well-documented reference materials, preferably certified reference materials. Where certified reference materials are not available, traceability to a definitive method should be established. Accreditation of the laboratory by the appropriate national accreditation scheme, which itself should conform to accepted standards[2], indicates that the laboratory is applying sound quality assurance principles. ISO Guide 25[3] describes the general guidelines for assessing a testing laboratory's technical competence. Although proficiency testing can be executed independently, accreditation assessments now use the information produced by proficiency testing[3].

Participation in proficiency testing schemes provides laboratories with an objective means of assessing and demonstrating the reliability of the data they are producing. Although there are several types of proficiency testing schemes, as described in ISO Guide 43[4], they all share the common feature of the comparison of test results obtained by one testing laboratory with those obtained by one or more other testing laboratories. Schemes may be "open" to any laboratory or participation may be by invitation only. Schemes may set out to assess the competence of laboratories undertaking a specific analysis in a specified matrix (i.e. lead in blood, fat in bonemeal) rather than the general type (food analysis) mentioned.

Although various protocols for the design and operation of proficiency testing schemes have been produced to cover particular areas of analytical chemistry, the need now is for a harmonised protocol for the organisation of proficiency testing schemes that would be universally accepted. The harmonised protocol detailed in Section 3 contains specific details and does not, therefore, coincide with ISO Guide 43[4]. In addition to describing the organisation and operation of the practical aspects of proficiency testing schemes, the document prescribes a minimal statistical treatment of the analytical data produced, which are primarily measurements of concentration.

Although various terms may be used to describe schemes conforming to this protocol (e.g. external quality assessment, performance schemes etc), the preferred term is "proficiency testing".

For any particular scheme the aims must be carefully described by the coordinating organisation. In addition, the coordinating organisation should appreciate that the procedure outlined below is to be regarded as the minimum that should be undertaken.

Schemes cannot cover all aspects of some areas of activity, and must be regarded as being representative of the particular sector of interest.

2. DEFINITIONS AND TERMINOLOGY USED IN PROTOCOL

2.1 Proficiency testing scheme

Methods of checking laboratory testing performance by means of interlaboratory tests.

[It includes comparison of a laboratory's results at intervals with those of other laboratories, with the main object being the establishment of trueness[5].]

2.2 Internal quality control (IQC)

The set of procedures undertaken by the laboratory for continuous monitoring of operations and results in order to decide whether the results are reliable enough to be released; IQC primarily monitors the batchwise accuracy of results on quality control materials, and precision on independent replicate analysis of test materials.

2.3 Quality assurance programme/system

The sum total of a laboratory's activities aimed at achieving the required standard of analysis. While IQC and proficiency testing are very important components a quality assurance programme must also include staff training, administrative procedures, management structure, auditing etc. Accreditation bodies judge laboratories on the basis of their quality assurance programme.

2.4 Testing laboratory

A laboratory that measures, examines, tests, calibrates or otherwise determines the characteristics or performance of materials or products.

2.5 Reference material (RM)

A material or substance one or more of whose properties values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to material[6].

Further information of the definition and use of reference material is available in the ISO REMCO documentation[7].

2.6 Certified reference material (CRM)

A reference material, accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes traceability to an accurate realisation of the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence[6].

Further information of the definition and use of certified reference materials is available in the ISO REMCO activities[7].

2.7 True value

The actual concentration of the analyte in the matrix.

2.8 Assigned value

The value to be used as the true value by the proficiency test coordinator in the statistical treatment of results. It is the best available estimate of the true value of the analyte in the matrix.

2.9 Target value for standard deviation

A numerical value for the standard deviation of a measurement result, which has been designated as a goal for measurement quality.

2.10 Interlaboratory test comparisons

Organisation, performance and evaluation of tests on the same items or materials on identical portions of an effectively homogeneous material, by two or more different laboratories in accordance with pre-determined conditions.

2.11 Coordinator

The organisation with responsibility for coordinating all of the activities involved in the operation of a proficiency testing scheme.

2.12 Accuracy

The closeness of agreement between a test result and the accepted reference value.

NOTE - The term accuracy, when applied to a set of test results, describes a combination of random components and a common systematic error or bias component.

2.13 Trueness

The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value.

NOTE - The measure of trueness is usually expressed in terms of bias.

2.14 Bias

The difference between the expectation of the test results and an accepted reference value.

NOTE - Bias is a systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the accepted reference value is reflected by a larger bias value.

2.15 Laboratory bias

The difference between the expectation of the test results from a particular laboratory and an accepted reference value.

2.16 Bias of the measurement method

The difference between the expectation of test results obtained from all laboratories using that method and an accepted reference value.

NOTE - One example of this in operation would be where a method purporting to measure the sulfur content of a compound consistently fails to extract all the sulfur, giving a negative bias to the measurement method. The bias of the measurement method is measured by the displacement of the average of results from a large number of different laboratories all using the same method. The bias of a measurement method may be different at different levels.

2.17 Laboratory component of bias

The difference between the laboratory bias and the bias of the measurement method.

NOTES

1. The laboratory component of bias is specific to a given laboratory and the conditions of measurement within the laboratory, and also it may be different levels of the test.

2. The laboratory component of bias is relative to the overall average result, not the true or reference value.

2.18 Precision

The closeness of agreement between independent test results obtained under prescribed conditions.

NOTES

1. Precision depends only on the distribution of random errors and does not relate to the accepted reference value.

2. The measure of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results. Higher imprecision is reflected by a larger standard deviation.

3. 'Independent test results' means results obtained in a manner not influenced by any previous result on the same or similar material.

3. ORGANISATION (PROTOCOL) OF PROFICIENCY TESTING SCHEMES

3.1 Framework

Test samples must be distributed on a regular basis to the participants who are required to return results within a given time. The results will be subject to statistical analysis by the coordinator and participants will be promptly notified of their performance. Advice will be available to poor performers and all participants will be kept fully informed of the progress of the scheme. Participants will be identified in Reports by code only.

The structure of the scheme for any one analyte or round in a series should be as follows:

1. coordinator organises preparation, homogeneity testing and validation of test material;

2. coordinator distributes test samples on a regular schedule;

3. participants analyse test portions and report results centrally;

4. results subjected to statistical analysis; performance of laboratories assessed;

5. participants notified of their performance;

6. advice available for poor performers, on request;

7. coordinator reviews performance of scheme;

8. next round commences.

Preparation for the next round of the scheme <u>may</u> have to be organised while the current round is taking place; details of the next round may have to be adjusted in the light of experience from the current round.

3.2 Organisation

Day to day running of the scheme will be the responsibility of the coordinator. The coordinator must document all practices and procedures in a quality manual (see Appendix I). Preparation of test materials will either be contracted-out or undertaken by the coordinator. The laboratory preparing the test material should have demonstrable experience in the area of analysis being tested. It is essential for the coordinator to retain control over the assessment of performance as this will help to maintain the credibility of the scheme. Overall direction of the scheme should be overseen by a small advisory panel having representatives (who should be practising laboratory scientists) from, for example, the coordinator, contract laboratories (if any), appropriate professional bodies, participants and end-users of analytical data.

3.3 Test materials

The test materials to be distributed in the scheme must be generally similar in type to the materials that are routinely analysed (in respect of composition of the matrix and the concentration range or quantity of the analyte). It is important that they are of acceptable homogeneity and stability. The assigned value will not be disclosed to the participants until after the results have been collated.

The bulk material prepared for the proficiency test must be sufficiently homogeneous for each analyte so that all laboratories will receive test samples that do not differ significantly in analyte concentration. The coordinator must clearly state the procedure used to establish the homogeneity of the test material. (Appendix II describes a recommended procedure). As a guide, the between-sample standard deviation should be less than 0.3 times the target value for the standard deviation.

Where possible the coordinating laboratory should also provide evidence that the test material is sufficiently stable to ensure that the material will not undergo any significant change throughout the duration of the proficiency test. Prior to distribution of the test samples, therefore, the stability of the matrix and the analytes it contains must be determined by carrying out analyses after it has been stored for an appropriate period of time. The storage conditions, most especially of time and temperature, used in the stability trials must represent those conditions likely to be encountered in the entire duration of the proficiency test. Stability trials must therefore take account of the transport of the test samples to participating laboratories as well as the conditions encountered purely in a laboratory environment. The concentrations of the various analytes must show no significant changes during the stability tests, the magnitude of a "significant change" being assessed from the knowledge of the variance expected for replicate analyses of the bulk material. When unstable analytes are to be assessed it may be necessary for the coordinating organisation to prescribe a date by which the analysis must be accomplished.

Ideally the quality checks on the samples referred to above should be performed by a different laboratory from that which prepared the sample, although it is recognised that this may cause difficulties to the coordinating organisation.

The number of test materials to be distributed per round will depend mainly on whether there is a requirement to cover a range of composition. Practical considerations will dictate an upper limit of six to the number of test materials per analyte.

Coordinators should consider any hazards that the test materials might pose and take appropriate action to advise any party that might be at risk (e.g. test material distributors, testing laboratories etc.) of the potential hazard involved.

3.4 Frequency of test sample distribution

The appropriate frequency for the distribution of test sample in any one series depends upon a number of factors of which the most important are:

i) the difficulty of executing effective analytical quality control;
ii) the laboratory throughput of test samples;
iii) the consistency of the results from previous rounds;
iv) the cost/benefit of the scheme;
v) the availability of suitable material for proficiency test schemes.

In practice the frequency will probably fall between once every two weeks and once every four months.

A frequency greater than once every two weeks could lead to problems in the turn round time of test samples and results. It might also encourage the belief that the proficiency testing scheme can be used as a substitute for internal quality control, an idea that is definitely to be discouraged. If the period between distributions extends much beyond four months, there will be unacceptable delays in identifying and correcting analytical problems, it could be difficult to monitor meaningful trends in a laboratory's performance and the impact of the scheme on the participants could be small.

There will be circumstances where consideration of the above factors may mean that it is acceptable to have a longer time scale between distribution of test samples. It would be one of the functions of the Advisory Panel to comment on the frequency of distribution appropriate for a particular scheme.

In addition this Panel would also provide advice on the areas to be covered in any particular sector of analytical chemistry. This is particularly difficult where there are a considerable number of diverse analyses in the sector.

3.5 Establishing the assigned value

The coordinator should give details on how the assigned value was obtained where possible with a statement of its traceability and its uncertainty.

There are a number of possible approaches to establishing the assigned value for the concentration of analyte and its uncertainty in a test material, but only four are normally considered.

3.5.1 Consensus value from expert laboratories

This value is the consensus of a group of expert laboratories that achieve agreement by the careful execution of recognised reference methods; it is the best procedure in most circumstances for determining the assigned value in representative materials. When such a value is used, the organising body should disclose the identities of the laboratories producing the individual results, the method of calculating the consensus value and, if possible, a statement of the traceability and of its uncertainty. The consensus value will normally be a robust mean[8] or the mode.

3.5.2 Formulation

This method comprises the addition of a known amount or concentration of analyte to a base material containing none. The method is especially valuable when it is the amount of analyte added to individual test portions that is subject to testing, as there is no requirement for ensuring a sufficiently homogeneous mixture in the bulk test material. In other circumstances problems might arise with the use of formulation, as follows.

(a) There is a need to ensure that the base material is effectively free from analyte or that the residual analyte concentration is accurately known,

(b) It may be difficult to mix the analyte homogeneously into the base material where this is required,

(c) The added analyte may be more loosely bonded than, or in a different chemical form from, that found in the typical materials that the test materials represent.

Unless these problems can be overcome, representative materials (containing the analyte in its normally occurring form in a typical matrix) are usually preferable. Where formulation is used, traceability to certified reference materials or reference methods should be cited if possible.

3.5.3 Direct comparison with certified reference materials

In this method, the test material is analysed along with appropriate certified reference materials by a suitable method under repeatability conditions. In effect the method is calibrated with the CRMs, providing direct traceability and an uncertainty for the value assigned to the test material. The CRMs must have both the appropriate matrix and an analyte concentration range that spans, or is close to, that of the test material. In some areas, the lack of CRMs will restrict the use of this method.

3.5.4 Consensus of participants

A value often advocated for the assigned value is the consensus (usually a robust mean or the mode) of the results of all of the participants in the round of the test. This value is clearly the cheapest and easiest to obtain. The method usually gives a serviceable value when the analysis is regarded as easy, for instance when a recognised method is applied to a major constituent. In an empirical method (where the method "defines" the content of the analyte), the consensus of a large number of laboratories can be safely regarded as the true value.

There are a number of drawbacks to the consensus of participants. At a fundamental level it is difficult to see how a traceability or an uncertainty could be attributed to such a value, unless all of the participants were using the same reference method. Other objections that can be levelled against the consensus are:
(a) there may be no real consensus amongst the participants and (b) the consensus may be biased by the general use of faulty methodology. Neither of these conditions is rare in the determination of trace constituents.

3.5.5 Choice between methods

The choice between these methods of evaluating the assigned value depends on circumstances and is the responsibility of the organising agency. It is usually advisable to have an estimate additional to the consensus of participants. Any significant deviations observed between the estimates must be carefully considered by the technical panel.

Empirical methods are used when the analyte is ill-defined chemically.In an empirical method, e.g. the determination of "fat", the true result (within the limits of measurement uncertainty) is produced by a correct execution of the method. It is clear that in these circumstances the analyte content is defined only if the method is simultaneously specified. Empirical methods can give rise to special problems in proficiency trials when a choice of such methods is available. If the assigned value is obtained from expert laboratories and the participants use a different empirical method, a bias may be apparent in the results even when no fault in execution is present. Likewise, if participants are free to choose between empirical methods, no valid consensus may be evident among them. Several recourses are available to overcome this problem:

(i) a separate value of the assigned value is produced for each empirical method used;

(ii) participants are instructed to use a prescribed method; or

(iii) participants are warned that a bias may result from using an empirical method different from that used to obtain the consensus.

3.6 Choice of analytical method

Participants will be able to use the analytical method of their choice except when otherwise instructed to adopt a specified method. Methods used must be validated by an appropriate means, e.g. collaborative trial, reference method etc. As a general principle, procedures used by laboratories participating in proficiency testing schemes should simulate those used in their routine analytical work.

Where an empirical method is used the assigned value will be calculated from results obtained using that defined procedure. If participants use a method which is not equivalent to the defining method, then an automatic bias in result must be expected when their performance is assessed (see Section 3.5.5).

3.7 Assessment of performance

Laboratories will be assessed on the difference between their result and the assigned value. A performance score will be calculated for each laboratory, using the statistical scheme detailed in Section 4.

3.8 Performance criteria

For each analyte in a round a criterion for the performance score may be set, where appropriate, against which the performance score obtained by a laboratory can be judged. A "running score" could be calculated to give an assessment of performance spread over a longer period of time; this would be based on results for several rounds.

The performance criterion will be set so as to ensure that the analytical data routinely produced by the laboratory is of a quality that is adequate for its intended purpose. It will not necessarily be appropriate to set the performance criterion at the highest level that the method is capable of providing.

3.9 Reporting of results

Reports issued to participants should be clear and comprehensive and include data on the distribution of results from all laboratories together with participant's performance score. The test results as used by the coordinator should be displayed also, to enable participants to check that their data have been correctly entered. Reports should be made available as quickly as possible after the return of results to the coordinating laboratory and, if at all possible, before the next distribution of samples.

Although ideally all results should be reported to participants, it may not be possible to achieve this in some very extensive schemes (e.g. where there are 700 participants each determining 20 analytes in any one round). Participants should, however, receive at least: (i) Reports in clear and simple format, and (ii) Results of all laboratories in graphical, e.g., histogram form.

3.10 Liaison with participants

Participants should be provided with a detailed information pack on joining the scheme. Communication with participants should be via a newsletter or annual report together with a periodic open meeting; participants should be advised immediately of any changes in scheme design or operation. Advice should be available to poor performers. Participants who consider that their performance assessment is in error must be able to refer the matter to the coordinator.

Feedback from laboratories should be encouraged, so that participants actively contribute to the development of the scheme. Participants should view it as <u>their</u> scheme rather than one imposed by a distant bureaucracy.

3.11 Collusion and falsification of results

Although proficiency testing schemes are intended primarily to help participants improve their analytical performance, there may be a tendency among some participants to provide a falsely optimistic impression of their capabilities. For example, collusion may take place between laboratories, so that truly independent data are not submitted. Laboratories may also give a false impression of their performance if they routinely carry out single analyses, but report the mean of replicate determinations on the proficiency test samples. Proficiency testing schemes should be designed to ensure that there is as little collusion and falsification as possible. For example, alternative samples could be distributed within one round, with no identifiable reuse of the materials in succeeding rounds. Also instructions to participants should make it clear that collusion is contrary to professional scientific conduct and serves only to nullify the benefits of proficiency testing to customers, accreditation bodies and analysts alike.

Although all reasonable measures should be taken by the coordinators to prevent collusion, it must be appreciated that it is the responsibility of the participating laboratories to avoid it.

3.12 Repeatability

Procedures used by laboratories participating in proficiency testing schemes should simulate those used in routine sample analysis. Thus, duplicate determinations on proficiency test samples should be carried out only if this is the norm for routine work. The result to be reported is in the same form (e.g., number of significant figures) as that normally reported to the customer. Some proficiency test coordinators like to include duplication in the tests to obtain a measure of repeatability proficiency. This should be allowed as a possibility in proficiency tests, but is not a requirement of this protocol.

4. GENERALISED STATISTICAL PROCEDURE FOR THE ANALYSIS OF RESULTS

The approach described here is intended to provide a transparent procedure by using accepted statistics without any arbitrary scaling factors.

4.1 Estimates of assigned value

The first stage in producing a score from a result x (a single measurement of analyte concentration (or amount) in a test material) is obtaining an estimate of the bias, which is defined as:

bias estimate = x − X

Where X is the true value.

In practice the assigned value, $\hat{X}$, which is the best estimate of X, is used.

Several methods are available for obtaining the assigned value (see Section 3.5).

If x is not a concentration measure, a preliminary transformation may be appropriate.

4.2 Formation of a z-score

Most proficiency testing schemes proceed by comparing the bias estimate (as defined above) with a target value for standard deviation that forms the criterion of performance. An obvious approach is to form the z-score given by

$$z = (x-\hat{X})/\sigma$$

where σ is the target value for standard deviation.

Although z has the form of a normal standard deviate there is no presumption that this necessarily will be the case. In some circumstances the technical panel may decide to use an estimate of the actual variation ($\tilde{s}$) encountered in a particular round of a trial in place of a target standard deviation. In that case $\tilde{s}$ should be estimated from the laboratories' results after outlier elimination, or by robust methods[8] for each analyte/material/round combination. A value of $\tilde{s}$ will thus vary from round to round. In consequence the z-score for a laboratory could not be compared directly from round to round. However the bias estimate ($x-\hat{X}$) for a single analyte/material combination could be usefully compared round by round for a laboratory, and the corresponding value of $\tilde{s}$ would indicate general improvement in "reproducibility" round by round.

A fixed value for σ is preferable and has the advantage that the z-scores derived from it can be compared from round to round to demonstrate general trends for a laboratory or a group of laboratories. It is suggested that whatever the value of σ is chosen it is a practical value and that it is accepted by participants. For some of the tests it is only necessary that the value chosen is sufficient clearly to discriminate in a pass/fail situation.

The value chosen can be arrived at in several ways:

4.2.1 By perception

The value of σ could be fixed arbitrarily, with a value based on a perception of how laboratories perform. The problem with this criterion is that both perceptions and laboratory performance may change with time. The value of σ therefore may need to be changed occasionally, disturbing the continuity of the scoring scheme. However, there is some evidence that laboratory performance responds favourably to a stepwise increase in performance requirements.

4.2.2 By prescription

The value of σ could be an estimate of the precision required for a specific task of data interpretation. This is the most satisfactory type of criterion, if it can be formulated, because it relates directly to the required information content of the data. Unless the concentration range is very small σ should be specified as a function of concentration.

This is frequently used in legislation where method performance characteristics may be specified.

4.2.3 By reference to validated methodology

Where a standard method is prescribed for the analysis, σ could be obtained by interpolation from the standard deviation of reproducibility obtained during appropriate collaborative trials.

4.2.4 By reference to a generalised model

The value of σ could be derived from a general model of precision, such as the "Horwitz Curve"(9). However, while this model provides a general picture of reproducibility, substantial deviation from it may be experienced for particular methods. It could be used if no specific information is available.

4.3 Interpretation of z-scores

If $\hat{X}$ and σ were good estimates of the population mean and standard deviation, and the underlying distribution were normal, then z would be approximately normally distributed with a mean of zero and a unit standard deviation. An analytical system can be described as "well behaved" when it complies with these conditions. Under these circumstances an absolute value of z ($|z|$) greater than three suggests poor performance.

Because z is standardised, it can be usefully compared between all analytes, test materials and analytical methods. Values of z obtained from diverse materials and concentration ranges can, therefore, with due caution (see Section 4.5), be combined to give a composite score for a laboratory in one round of a proficiency test. Moreover the meaning of z-scores can be immediately appreciated, i.e. values of $|z|<2$ would be very common and values of $|z|>3$ would be very rare in well behaved systems.

Schemes explicitly based on the z-score method include the "Laboratory Accreditation and Audit Protocol"(10). The z-score method is also implicit in the modified "variance index" method of Whitehead et al(11), where scaling to a "chosen coefficient of variation" (i.e. relative standard deviation) effectively gives a z-value multiplied by an arbitrary factor.

4.4 An alternative score

An alternative type of scoring, here called Q-scoring, is based not on the standardised value but on the relative bias, namely

$$Q = (x-\hat{X})/\hat{X}$$

where x and $\hat{X}$ have their previous meaning.

Although not recommended in this protocol a number of sectors, for example occupational hygiene, use this type of approach.

The scoring does have the disadvantage that the significance of any result is not immediately apparent.

The alternative type of scoring is described in greater detail in Appendix V.

4.5 Combination of results of a laboratory within a round of the trial

It is common for several different analyses to be required within each round of a proficiency test. While each individual test furnishes useful information, many participants want a single figure of merit that will summarise the overall performance of the laboratory within a round. This approach may be appropriate for the assessment of long term trends. However, there is a danger that such a combination score will be misinterpreted or abused by non-experts, especially outside the context of the individual scores. Therefore the general use of combination scores is not recommended, but it is recognised that they may have specific applications if based on sound statistical principles and used with due caution.

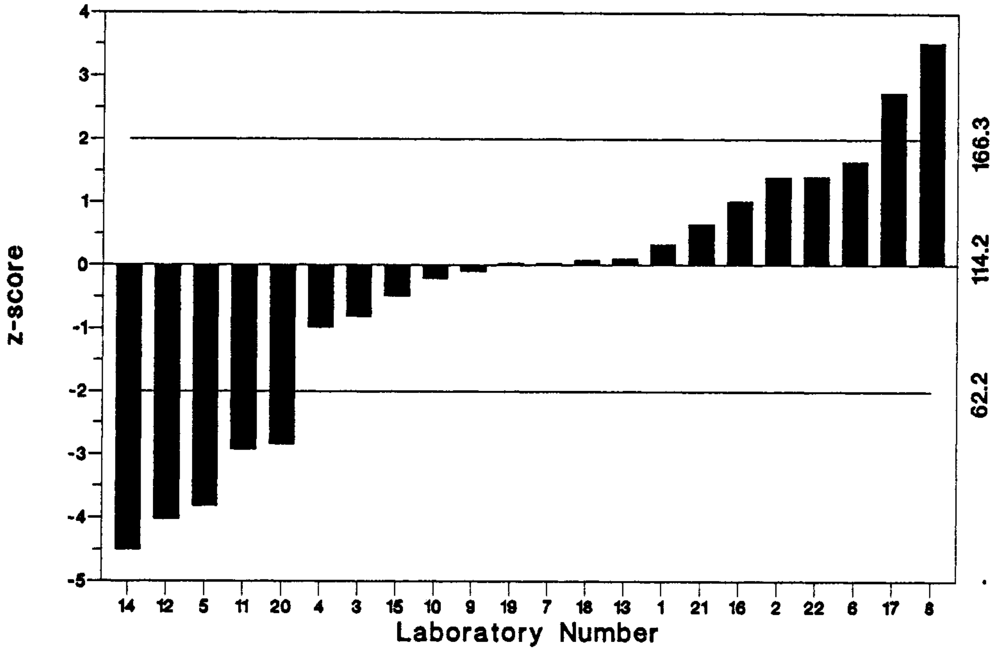

FIGURE 1: Report 0502
Z-SCORE FOR HEXACHLOROBENZENE IN OIL (114.2ug/kg)

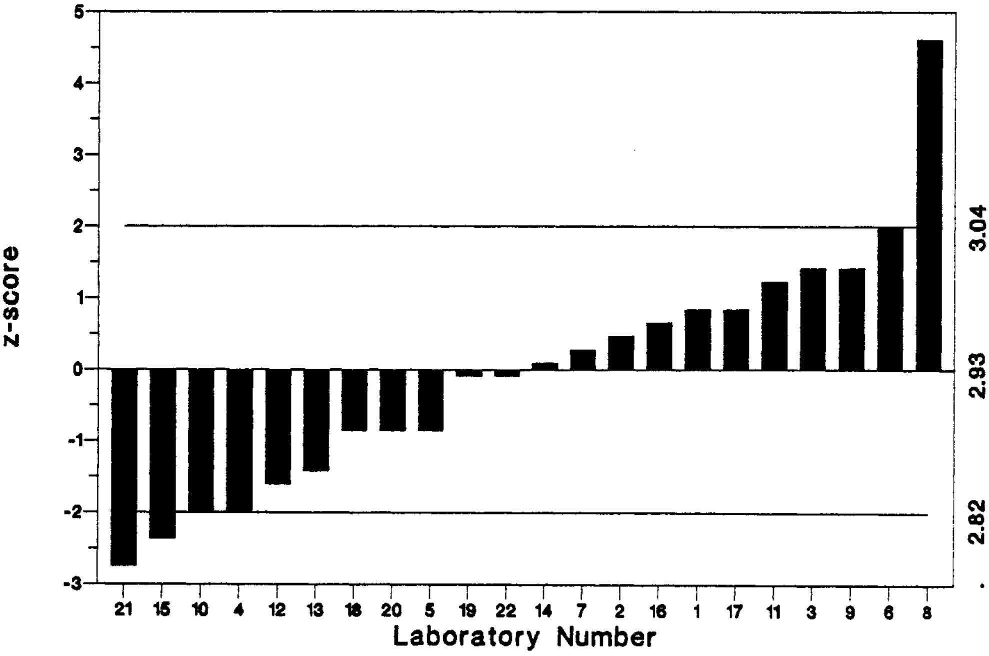

FIGURE 2: Report 0603
Z-SCORE FOR NITROGEN IN CEREAL (2.93g/100g)

It is especially emphasised that there are limitations and weaknesses in any scheme that combines z-scores from dissimilar analyses. If a single score out of several produced by a laboratory were outlying, the combined score may well be not outlying. In some respects this is a useful feature, in that a lapse in a single analysis is downweighted in the combined score. However, there is a danger that a laboratory may be consistently at fault only in a particular analysis, and frequently report an unacceptable value for that analysis in successive rounds of the trial. This factor may well be obscured by the combination of scores.

The procedures that may be used are described in Appendix III.

4.6 Running scores

While the combination scores discussed above and detailed in Appendix IV give a numerical account of the performance of a laboratory in a single round of the proficiency test, for some purposes it may be useful to have a more general indicator of the performance of a laboratory over time.

While the value of such indicators is questionable they can be constructed simply and give a smoothed impression of the scores over several rounds of the test.

Some procedures that may be used are described in Appendix IV. It must be stressed that, as with combination scores (see Section 4.5), it is difficult to produce running scores that are not prone to misinterpretation etc.

4.7 Classification, ranking and other assessment of proficiency data

Classification is not the primary aim of proficiency testing. However, it is possible that accreditation agencies will use proficiency test results for this purpose, so it is essential that any classification used should be statistically well-founded.

4.7.1 Classification

If the frequency distribution of a proficiency score is known or presumed, then significance can be attributed to results according to the quantiles of that distribution. In a well-behaved analytical system z-scores would be expected to fall outside the range $-2<z<2$ in about 5% of instances, and outside the range $-3<z<3$ only in about 0.3%. In the latter case it could be interpreted that the probability is so small for a "well-behaved" system, that it almost certainly represents a poor performance. It would therefore be possible to classify scores as:

$\|z\|\leq 2$	Satisfactory
$2<\|z\|<3$	Questionable
$\|z\|\geq 3$	Unsatisfactory

z scores are broadly comparable but the use of any classification must, in practice, be treated with care as the knowledge of the relevant probabilities rests on assumptions that may not be fulfilled: (i) that the appropriate values of $\hat{X}$ and σ have been used; and (ii) that the underlying distribution of analytical errors is normal, apart from outliers. In addition the division of a continuous measure into a few named classes has little to commend it from the scientific point of view, although it may have a psychological effect on the participants. Consequently, classification is not recommended in proficiency tests. "Decision limits" based on z-scores may be used as an alternative where necessary.

4.7.2 Ranking

Laboratories participating in a round of a proficiency trial are sometimes ranked on their combined score for the round or on a running score. Such a ranked list is used for encouraging better performance in poorly ranked laboratories by providing a comparison among the participants. However, ranking is not recommended as it is an inefficient use of the information available and may be open to mis-interpretation. A histogram is a more effective method of presenting the same data.

5. AN OUTLINE OF HOW ASSIGNED VALUES AND TARGET VALUES MAY BE SPECIFIED AND USED

A hypothetical example of how assigned values and target values may be specified and used is given in Appendix VI.

6. REFERENCES

The references cited throughout this document and its Appendices are given below. Additional references are given in ISO Guide 43.

1. "Protocol for the Design, Conduct and Interpretation of Collaborative Studies", Edited W Horwitz, Pure & Appl. Chem. 1988, 60 (6), 855-864.

2. "Testing Laboratory Accreditation Systems - General Recommendations for the Acceptance of Accreditation Bodies", ISO Guide 54, 1988, Geneva.

3. "General Requirements for the Competence of Calibration and Testing Laboratories", ISO Guide 25, 3rd Edition 1990, Geneva.

4. "Development and Operation of Laboratory Proficiency Testing", ISO Guide 43, 1984, Geneva.

5. "Accuracy (Trueness and Precision) of Measurement Methods and Results — Part 1: General Principles and Definitions", ISO DIS-5725-1, Geneva.

6. "Terms and Definitions used in Connection with Reference Materials", ISO Guide 30, 1981, Geneva.

7. "Uses of Certified Reference Materials", ISO Guide 33, 1989, Geneva.

8. Analytical Methods Committee Report "Robust Statistics, Part 1". Analyst 1989, 114, 1693-1697.

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

10. "Laboratory Accreditation and Audit Protocol", Food Inspection Directorate, Food Production and Inspection Branch, Agriculture Canada, April 1986.

11. "A comparative survey of the results of analyses of blood serum in clinical chemistry laboratories in the United Kingdom", Whitehead, T.P., Browning, D. M. and Gregory, A., J. Clin. Pathol., 1973, 26, 435-445.

APPENDIX I: SUGGESTED HEADINGS IN A QUALITY MANUAL FOR ORGANISATION OF PROFICIENCY TESTING SCHEMES (NOT NECESSARILY IN THIS ORDER)

1. Quality policy
2. Organisation of agency
3. Staff, including responsibilities
4. Documentation control
5. Audit and review procedures
6. Aims, scope, statistical design and format (including frequency) of proficiency testing programmes
7. Procedures covering - sample preparation
   - testing of sample homogeneity
   - equipment
   - suppliers
   - logistics (e.g. sample dispatch)
   - analysis of data

8. Preparation and issuing of report.
9. Action and feedback by participants when required
10. Documentation of records for each programme
11. Complaints handling procedures
12. Policies on confidentiality and ethical considerations
13. Computing information, including maintenance of hardware and software
14. Safety and other environmental factors
15. Sub-contracting
16. Fees for participation
17. Scope of availability of programme to others

APPENDIX II: A RECOMMENDED PROCEDURE FOR TESTING A MATERIAL FOR SUFFICIENT HOMOGENEITY

The procedure to be followed by the Laboratory preparing proficiency test materials is as follows:

1. Prepare the whole of the bulk material in a form that is thought to be homogeneous, by an appropriate method.

2. Divide the material into the containers that will be used for dispatch to the participants.

3. Select a minimum (n) of 10 containers strictly at random.

4. Separately homogenise the contents of each of the n selected containers and take two test portions.

5. Analyse the 2n test portions in a random order under repeatability conditions by an appropriate method. The analytical method used must be sufficiently precise to allow a satisfactory estimation of S_s.

6. Form an estimate (S_s^2) of the sampling variance and an estimate (S_a^2) of the analytical variance by one-way analysis of variance, without exclusion of outliers.

7. Report values of $\bar{x}$, S_s, S_a, n and the result of the F-test.

8. If σ is the target value for standard deviation for the proficiency test at analyte concentration = $\bar{x}$, the value of S_s/σ should be less than 0.3 for sufficient homogeneity.

EXAMPLE – Copper in Soya Flour ($\mu g\ g^{-1}$)

Sample No	Copper Content	
	1	2
1	10.5	10.4
2	9.6	9.5
3	10.4	9.9
4	9.5	9.9
5	10.0	9.7
6	9.6	10.1
7	9.8	10.4
8	9.8	10.2
9	10.8	10.7
10	10.2	10.0
11	9.8	9.5
12	10.2	10.0

Grand mean = 10.02

Analysis of Variance

Source of variation	df	Sum of Squares	Mean Square	F
Between samples	11	2.54458	0.231326	3.78
Analytical	12	0.735000	0.06125	

Critical value of F ($p=0.05$, $v_1=11$, $v_2=12$) is 2.72 < 3.78

There are significant differences between samples

$S_a = \sqrt{}\ 0.0613 = 0.25$

$S_s = ((0.2313 - 0.0613)/2)^{1/2} = 0.29$

$\sigma =$ 1.1 (This is an example value of a target value for reference standard deviation and is not derived from the data)

$S_s/\sigma = 0.29 / 1.1 = 0.26 < 0.3$

Although there are significant differences between samples (F - test), the material is sufficiently homogeneous for the purpose of the proficiency trial, as $S_s/\sigma = 0.26$ is less than the maximum recommended value of 0.3.

APPENDIX III: COMBINATION OF RESULTS OF A LABORATORY WITHIN ONE ROUND OF A TRIAL

The general use of combination scores is not recommended, but it is recognised that they may have specific applications if used with due caution.

1. Introduction

Several methods of combining independent z-scores produced by a laboratory in one round of the test seem potentially appropriate; for example:

(i) The sum of scores, $SZ = \Sigma z$

(ii) The sum of squared scores, $SSZ = \Sigma z^2$

(ii) The sum of absolute values of the scores, $SAZ = \Sigma\ |z|$

These statistics fall into two classes. The first class (containing only SZ) uses information about the signs of the z-scores, while the alternative class (SSZ and SAZ) provides information about only the size of scores, i.e. the magnitude of biases. Of the latter, the sum of the squares is more tractable mathematically and is therefore the preferred statistic, although it is rather sensitive to single outliers. SAZ may be especially useful if there are extreme outliers or many outlying laboratories, but its distribution is complicated and its use is not, therefore, recommended.

2. Sum of scores, SZ

The distribution of SZ is zero-centred with variance m, where m is the number of scores being combined. Thus SZ could not be interpreted on the same scale as the z scores. However, a simple scaling restores the unit variance, giving a rescaled sum of scores $RSZ = \Sigma z/\sqrt{m}$ which harmonises the scaling. In other words, both z and RSZ can be interpreted as standard normal deviates.

SZ and RSZ have the advantage of using the information in the signs of the biases. Thus if a set of z-scores were (1.5, 1.5, 1.5, 1.5) the individual results would be regarded as non-significant positive scores. However, regarded as a group, the joint probability of observing four such deviations together would be small. This is reflected in the RSZ value of 3.0, which indicates a significant event. This information would be useful in detecting a small consistent bias in an analytical system, but would not be useful in combining results from several different systems, where a consistent bias would not be expected, and is unlikely to be meaningful.

Another feature of the RSZ is the tendency for errors of opposite sign to cancel. In a well-behaved situation (i.e. when the laboratory is performing without bias according to the designated σ value) this causes no problems. If the laboratory were badly behaved, however, the possibility arises of the fortuitous cancellation of significantly large z values. Such an occurrence would be very rare by chance.

These restrictions on the use of RSZ serve to emphasise the problems of using combination scores derived from various analytical tests. When such a score is used, it should be considered simultaneously with the individual scores.

3. Sum of squared scores, SSZ

This combination score has a chi-squared (χ^2) distribution with m degrees of freedom for a well-behaved laboratory. Hence there is no simple possibility for interpreting the score on a common scale with the z-scores. However, the quantiles of the χ^2 distribution can be found in most compilations of statistical tables.

SSZ takes no account of the signs of the z-values, because of the squared terms. Thus, in the example considered previously, where the z-scores are (1.5, 1.5, 1.5, 1.5), we find SSZ = 9.0, a value that is not significant at the 5% level, and does not draw enough attention to the unusual nature of the results as a group. However, in proficiency tests, we are concerned much more with the magnitude of deviations than with their direction, so SSZ seems appropriate for this use. Moreover, the problem of chance cancellation of significant z-scores of opposite sign is eliminated. Thus the SSZ has advantages as a combination score for diverse analytical tests, and is to an extent complementary to RSZ. A related score, (SSZ/m), is used in the "Laboratory Audit and Accreditation Scheme".

APPENDIX IV: CALCULATION OF RUNNING SCORES

The general use of running scores is not recommended, but it is recognised that they may have specific applications if used with due caution. The usual procedure for calculating running scores is to form a "moving window" average. The procedure can be applied to z or a combination score.

As an example, a running Ƶ - score covering the current (n-th) round and the previous k rounds could be constructed as follows:

$$\text{RƵ}_n = \sum_{j=n-k}^{n} z_j \;/\; (k+1)$$

where z_j is the z-score for the material in the j-th round.

The running score has the alleged advantage that instances of poor performance restricted to one round are smoothed out somewhat, allowing an overall appraisal of performance. On the other hand, an isolated serious deviation will have a "memory effect" in a simple moving window average that will persist until (k+1) more rounds of the trial have passed. This might have the effect of causing a laboratory persistently to fail a test on the basis of the running score, long after the problem has been rectified.

Two strategies for avoiding undue emphasis on an isolated bad round are suggested. Firstly, individual or combined scores can be restrained within certain limits. For example, we could apply a rule such as:

if $|z|>3$ then $z' = \pm 3$, the sign being the same as that of z

where z is the raw value of a z-score, and the modified value z' is limited to the range ± 3.

The actual limit used could be set in such a way that an isolated event does not raise the running score above a critical decision level for otherwise well behaved conditions.

As a second strategy for avoiding memory effects, the scores could be "filtered" so that results from rounds further in the past would have a smaller effect on the running score. For example, exponential smoothing uses:

$$\hat{z}_n = \sum_{i=0}^{\infty} \alpha^i z_{n-i}/(1-\alpha)$$

calculated by:

$$\hat{z}_n = (1 - \alpha)z_n + \alpha\hat{z}_{n-1}$$

where α is a parameter between zero and one, controlling the degree of smoothing.

APPENDIX V: AN ALTERNATIVE SCORING PROCEDURE FOR PROFICIENCY TESTING SCHEMES

An alternative type of scoring, which will be called Q-scoring, is based not on the standardised value but on the relative bias, namely

$$Q = (x-\hat{X})/\hat{X}$$

where x and $\hat{X}$ have their previous meaning. This type of score relates directly to the analytical errors, without any reference to a value of σ which would need to come either from the participants' data or from an imposed performance standard.

It would be expected that the overall distribution of Q would be centred on zero. This must be so where all-participant means are used as the estimate of the true value, provided that the number of outliers is relatively low. It should also be so where expert laboratory consensus means are used, as long as the group of non-expert laboratories does not show an overall bias relative to the experts. When the true value is defined as a known addition, the Q distribution will be centred on zero provided that this true value is correct, and that there is no widespread use of methodology leading to biased results. In many cases, the actual distribution of Q-scores can be used to test the underlying assumptions.

The distribution of Q-scores cannot be predicted. Organisers of a scheme will need to examine the distribution of scores when laying down criteria for assessing whether or not performance is acceptable. In practice, the distribution has often been found to be close to normal.

An advantage of Q-scoring is that it gives a direct measure of the error associated with a determination. This can subsequently be compared with a performance standard which is judged to be appropriate to the purpose of the determination[1]. If different end-uses of the determination require different performance standards, the Q-score can be used in comparison with whichever standard is most appropriate. Furthermore, if the organiser of a scheme decide at any time that a change in performance standard is justified, previously generated results can easily be compared retrospectively with a revised standard.

Reference

1. Initial Experience with the Workplace Analysis Scheme for Proficiency (WASP), H M Jackson and N G West, Annals of Occupational Hygiene, in press.

APPENDIX VI: AN OUTLINE EXAMPLE OF HOW ASSIGNED VALUES AND TARGET VALUES MAY BE SPECIFIED AND USED

This is intended to be an example of how assigned values and target values may be calculated and used according to the protocol. Numerical details have been specified for the purposes of illustration only; real schemes will have to take account of factors specific to their area.

1. Scheme

The example requires that there will be four distributions of materials per year, dispatched by post on the Monday of the first full working week of January, April, July and October. Results must reach the organisers by the last day of the respective month. A statistical analysis of the results will be dispatched to participants within two weeks of the closing dates. This example considers the results from one particular circulation of two test materials for the determination of two analytes.

2. Testing for Sufficient Homogeneity

In accordance with the procedure described in Appendix II, where a full example is given.

3. Analyses required

The analyses required in each round will be:

(i) Hexachlorobenzene in an oil;

(ii) Kjeldahl nitrogen in a cereal product.

4. Methods of Analysis and Reporting of Results

No method is specified, but the target values were determined using a standard method, and participants must provide an outline of the method actually used, or give a reference to a documented method.

Participants must report a single result, in the same form as would be provided for a client.

Individual reported values are given in Table 1.

5. Assigned Values

5.1 Hexachlorobenzene in Oil

Take the estimate of assigned analyte concentration $\hat{X}$ for the batch of material as the robust mean of the results of six expert laboratories.

Reference Laboratory	Result (µg/kg)
7	115.0
9	112.0
10	109.0
13	117.0
18	116.2
19	115.0

$\hat{X}$ is 114.23 µg/kg: traceability was obtained using a reference method calibrated using in house reference standards and the uncertainty on the assigned value was determined to be ±10µg/kg from a detailed assessment of this method by the reference laboratories.

5.2 Kjeldahl Nitrogen in a Cereal Product

Take the assigned value of analyte concentration $\hat{X}$ for the batch of material as the median of the results from all laboratories.

Table 1 TABULAR EXAMPLES

Analyte	Hexachlorobenzene in Oil		Nitrogen in Cereal	
Assigned Value	114.2µg/kg		2.93g/100g	
Laboratory	Result	z-Score	Result	z-Score
001	122.6	0.3	2.97	0.9
002	149.8	1.4	2.95	0.5
003	93.4	-0.8	3.00	1.4
004	89.0	-1.0	2.82	-2.0
005	17.4	-3.8	2.88	-0.9
006	156.0	1.7	3.03	2.0
007	115.0	0.0	2.94	0.3
008	203.8	3.5	3.17	4.7
009	112.0	-0.1	3.00	1.4
010	109.0	-0.2	2.82	-2.0
011	40.0	-2.9	2.99	1.2
012	12.0	-4.0	2.84	-1.6
013	117.0	0.1	2.85	-1.4
014	0.0	-4.5	2.93	0.1
015	101.8	-0.5	2.80	-2.4
016	140.0	1.0	2.96	0.7
017	183.5	2.7	2.97	0.9
018	116.2	0.1	2.88	-0.9
019	115.0	0.0	2.92	-0.1
020	42.3	-2.8	2.88	-0.9
021	130.8	0.7	2.78	-2.8
022	150.0	1.4	2.92	-0.1

6. Target Values for Standard Deviation

6.1 Hexachlorobenzene in Oil

In the example used in this Appendix the %RSD_R value has been calculated from the Horwitz Equation

(RSD_R in % = $2^{(1-0.5\log \hat{X})}$).

The target value for the standard deviation (σ) will therefore be:

$\sigma_1 = 0.222\hat{X}$ µg/kg

6.2 Kjeldahl Nitrogen in a Cereal Product

In the examples used in this Appendix the %RSD_R value has been calculated from published collaborative trials.

The target value for the standard deviation (σ) is given by:

σ_2 = 0.018$\hat{X}$ g/100g

7. Statistical Analysis of Results of the Test

7.1 Hexachlorobenzene in Oil: Formation of z-Score

Calculate:

$$z = (x - \hat{X}) / \sigma$$

for each individual result (x) using the values of $\hat{X}$ and σ derived above. These results are shown in Table 1.

7.2 Kjeldahl Nitrogen in a Cereal Product: Formation of z-Score

Calculate:

$$z = (x - \hat{X}) / \sigma$$

for each individual result (x) using the values of $\hat{X}$ and σ derived above. These results are shown in Table 1.

8. Display of Results

8.1 z Scores Tables

The individual results for hexachlorobenzene pesticide in oil and for Kjeldahl nitrogen in a cereal product, together with associated z-scores are displayed in Tabular form (Table 1).

8.2 Histograms for z-Scores

The z-scores for hexachlorobenzene pesticide in oil and for Kjeldahl nitrogen in a cereal product are also displayed as bar-charts; figure 1 and 2 respectively.

9. Decision Limits

Results with an absolute value for z of less than 2 will be regarded as satisfactory.

Remedial action will be recommended when any of the z-scores exceed an absolute value of 3.0.

In this example, such results are:

Laboratories 005, 008, 012, 014 for hexachlorobenzene pesticide in oil,

and

Laboratory 008 for Kjeldahl nitrogen in a cereal product.