Pure Appl. Chem., Vol. 78, No. 1, pp. 145–196, 2006.
doi:10.1351/pac200678010145
© 2006 IUPAC

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

ANALYTICAL CHEMISTRY DIVISION*

INTERDIVISIONAL WORKING PARTY FOR HARMONIZATION OF QUALITY ASSURANCE SCHEMES

THE INTERNATIONAL HARMONIZED PROTOCOL FOR THE PROFICIENCY TESTING OF ANALYTICAL CHEMISTRY LABORATORIES

(IUPAC Technical Report)

Prepared for publication by
MICHAEL THOMPSON[1], STEPHEN L. R. ELLISON[2,‡], AND ROGER WOOD[3]

[1]School of Biological and Chemical Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK; [2]LGC Limited, Queens Road, Teddington Middlesex, TW11 0LY, UK; [3]Food Standards Agency, c/o Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

*Membership of the Analytical Chemistry Division during the final preparation of this report was as follows:

President: K. J. Powell (New Zealand); ***Titular Members***: D. Moore (USA); R. Lobinski (France); R. M. Smith (UK); M. Bonardi (Italy); A. Fajgelj (Slovenia); B. Hibbert (Australia); J.-Å. Jönsson (Sweden); K. Matsumoto (Japan); E. A. G. Zagatto (Brazil); ***Associate Members***: Z. Chai (China); H. Gamsjäger (Austria); D. W. Kutner (Poland); K. Murray (USA); Y. Umezawa (Japan); Y. Vlasov (Russia); ***National Representatives***: J. Arunachalam (India); C. Balarew (Bulgaria); D. A. Batistoni (Argentina); K. Danzer (Germany); E. Domínguez (Spain); W. Lund (Norway); Z. Mester (Canada); ***Provisional Member***: N. Torto (Botswana).

‡Corresponding author: E-mail: s.ellison@lgc.co.uk

The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories

(IUPAC Technical Report)

Abstract: The international standardizing organizations—AOAC International, ISO, and IUPAC—cooperated to produce the International Harmonized Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories. The Working Group that produced the protocol agreed to revise that Protocol in the light of recent developments and the experience gained since it was first published. This revision has been prepared and agreed upon in the light of comments received following open consultation.

Keywords: harmonized; IUPAC Analytical Chemistry Division; uncertainty; analysis; proficiency testing; protocol.

CONTENTS

PART 1: FOREWORD AND INTRODUCTION

1.0 Foreword

In the 10 years since the first version of this protocol was published [1], proficiency testing has burgeoned. The method has become widely used in many sectors of chemical analysis, and many new proficiency testing schemes have been launched worldwide [2]. A detailed study of proficiency testing for the analytical chemistry laboratory has been undertaken [3]. The International Organization for Standardization (ISO) has published a guide to proficiency testing [4] and a standard on statistical methods for use in proficiency testing [5]. The International Laboratory Accreditation Corporation (ILAC) has published a document on the quality requirements for proficiency testing [6], and many proficiency testing schemes have now been accredited. In addition, the clarification over the last decade of the application of the uncertainty concept to chemical measurement has had an effect on the way in which we view proficiency testing. This extraordinary development of proficiency testing is both a recognition of its unrivalled power to expose unexpected problems in analysis and the current requirement of participation in a proficiency testing scheme as part of the accreditation of analytical laboratories.

As a result of all this activity in many different analytical sectors, together with the considerable amount of research that has been conducted, the analytical community has built up a large body of new experience with proficiency testing. It is pleasing to note that no substantive modifications of the fundamental ideas and principles of the 1993 Harmonized Protocol [1] are required to accommodate this new experience. However, the additional experience shows a need, and provides a basis, for refinement of our approach to many aspects of proficiency testing and for more specific and definite recommendations in some areas. Further, the original Harmonized Protocol was largely concerned with the organization of proficiency testing schemes, and is therefore addressed mainly to providers of schemes. The increased importance of proficiency testing scheme data has, however, generated a need for additional guidance on the interpretation of results of schemes by both scheme participants and "end-users" of analytical data (such as laboratory customers, regulators, and other stakeholders in laboratory quality). All these factors call for an update of the 1993 Harmonized Protocol.

The revision also provides an opportunity to point out that some important aspects of proficiency testing remain incompletely documented at this time. In addition, we must recognize that the variety of possible approaches featured in the ISO documents is intended to be comprehensive and to cover all fields of measurement. Practical experience in chemical analysis strongly suggests that a restricted subset from this wide range of approaches provides an optimal approach for routine analytical work. This updated Harmonized Protocol is, therefore, not merely a collation of extracts from other documents, but an optimal subset of methods, based on detailed practical experience of managing proficiency testing schemes, interpreted specifically for analytical chemistry, and incorporating the newest ideas.

Producing an updated Protocol further allows us to emphasize the importance of professional judgement and experience both for the provision of proficiency testing schemes and for the participants in acting appropriately upon the results. Adherence to a protocol obviously implies that certain actions *must* be carried out. But the means by which they are carried out needs to be contingent to some extent on the particular application, and the range of possible applications is both large and changing with time. Further, any experienced analytical practitioner will readily recognize that real-life test materials and methods in the rapidly changing field of analytical chemistry will invariably generate occasional unexpected behavior that demands expert consideration and vigilance. Thus, we regard it as unsafe to exclude all scope for expert judgement and to replace it with inflexible rules. The structure of this document reflects that philosophy. The Protocol proper comes first and comprises a series of relatively short sections outlining the essential actions required for schemes claiming to adhere to it. This is followed by a number of longer sections and appendices that discuss the options available for carrying out the Protocol and the reasons why particular recommendations are made. The appendices include independent sections specifically for participants and for end-users of data to assist them with interpretation of proficiency testing scheme data.

Finally we note that, although this document retains the title of "Protocol", we are not advocating the philosophy that there is only one good way to conduct proficiency testing. This document simply outlines what has been shown to be good and effective practice for analytical chemists in the majority of instances. We recognize that circumstances may dictate that alternative procedures may be required, and that such choices are the responsibility of the provider with the help of the scheme's advisory committee. We also point out that this protocol is largely restricted to the scientific and technical aspects of proficiency testing used as a tool for improving measurement performance. It does not, therefore, address issues such as qualification or disqualification of laboratories or staff for particular purposes, or the accreditation of proficiency testing providers or specific schemes.

1.1 Rationale for proficiency testing

For a laboratory to produce consistently reliable data, it must implement an appropriate program of quality-assurance and performance-monitoring procedures. Proficiency testing is one of these procedures. The usual format for proficiency testing schemes in analytical chemistry is based on the distribution of samples of a test material to the participants. Participating laboratories ("participants") generally know the test material has been sent by a proficiency scheme provider, but occasionally the material may be received "blind" (i.e., it is received from a normal customer of the laboratory). The participants analyze the material without knowledge of the correct result and return the result of the measurement to the scheme provider. The provider converts the results into scores that reflect the performance of the participant laboratory. This alerts the participant to unexpected problems that might be present, and spurs the management to take whatever remedial action is necessary.

The ethos of this Harmonized Protocol is that proficiency testing should provide information on the fitness-for-purpose of analytical results provided by participants, to assist them in meeting requirements. This can be achieved when

- criteria for assessing results take fitness-for-purpose into account, so that scores inform participants when they need to improve their performance to satisfy customer (or stakeholder) needs;
- the circumstances of proficiency testing are close to those prevailing during routine analysis, so that the outcome represents "real life"; and
- the method of scoring should be simple, and where at all possible, consistent over the whole realm of analytical measurement, to ensure ready interpretation by participants and customers.

While the first consideration of proficiency testing is to provide a basis for self-help for each participant, it would be disingenuous to ignore the fact that other uses are made of proficiency testing results. Participants commonly use their scores to demonstrate competence to potential customers and accreditation assessors, and this has the unfortunate effect of pressurizing analysts to excel in the proficiency tests rather than simply to assess routine procedures. Participants should make every effort to avoid such a tendency as, for the most part, it is impossible for scheme providers to detect or eliminate it. Participants must also be diligent in avoiding any misinterpretation of accumulated scores.

1.2 Proficiency testing in relation to other quality-assurance methods

A comprehensive scheme of quality assurance (QA) in analytical chemistry laboratories would include the following elements in addition to proficiency testing: the validation of analytical methods [7]; the use of certified reference materials (CRMs) (where available) [8]; and the employment of routine internal quality control (IQC) [9]. Traditionally, the validation of an analytical method implies that its performance characteristics—trueness, precision under various conditions, and calibration linearity and so on—are known sufficiently well. In more modern terms, that means that we have estimated the method's measurement uncertainty, in a one-off operation under plausible conditions of use, and found it to be potentially fit for purpose. Method validation ideally involves *inter alia* the use of matrix-appropriate CRMs, if they are available, for calibration or for checking existing calibrations if matrix effects are present. Where CRMs are not available, other expedients have to be employed.

IQC should be conducted as a matter of routine and involves the analysis of one or more "control materials" within every run of analysis. This procedure, combined with the use of control charts, ensures that the factors determining the magnitude of the uncertainty have not changed materially since fitness-for-purpose was originally demonstrated in the validation process. In other words, the uncertainty estimated at validation is demonstrated (within the limits of statistical variation) to apply to each individual run executed subsequently. Preparation of a control material also ideally involves the use of CRMs to establish traceability of the measurand values assigned to it.

In principle, method validation and IQC alone are sufficient to ensure accuracy. In practice, they are often less than perfect. Proficiency testing is, therefore, the means of ensuring that these two within-laboratory procedures are working satisfactorily. In method validation, unknown influences may interfere with the measurement process and, in many sectors, CRMs are not available. Under such conditions, traceability is hard to establish, and unrecognized sources of error may be present in the measurement process. Laboratories with no external reference could operate for long periods with biases or random variations of serious magnitude. Proficiency testing is a means of detecting and initiating the remediation of such problems (see Appendix 6). Its main virtue is that it provides a means by which participants can obtain an external and independent assessment of the accuracy of their results.

PART 2: THE HARMONIZED PROTOCOL; ORGANIZATION OF PROFICIENCY TESTING SCHEMES

2.1 Scope and field of application

This protocol is applicable where

- the principal aim is the assessment of laboratory performance against established criteria based on fitness for a common purpose;
- compliance with these criteria may be judged on the basis of the deviation of measurement results from assigned values; and
- participants' results are reported on an interval scale or a ratio scale.

Note: These conditions apply widely in the assessment of analytical chemistry laboratories performing routine testing, but also in many other fields of measurement and testing.

This protocol is not intended for the assessment of calibration services and therefore makes no provision for the use by the scheme provider of uncertainty information provided with participant results. Nor does it provide criteria for the assessment, certification, or accreditation of proficiency scheme providers.

2.2 Terminology

Technical words in this document follow their ISO definitions, where such are available. Abbreviations follow the IUPAC Compendium of Analytical Nomenclature (1997). The following additional terms occur frequently in this document:

- *Proficiency testing provider* (*"the scheme provider" or "provider"*): Organization responsible for the coordination of a particular proficiency scheme.
- *(Proficiency) test material*: The material that is distributed for analysis by participants in a proficiency test.

- *Distribution unit*: A packaged portion of the test material that is sent or ready to be sent to a participant laboratory.
- *Test portion*: The part of the distribution unit that is used for analysis.
  Note: The test portion may comprise an entire distribution unit or portion thereof.
- *Series*: Part of a proficiency scheme, defined by a particular range of test materials, analytes, analytical methods, or other common features.
- *Round*: A single distribution episode in a series.

2.3 Framework of proficiency testing

2.3.1 Scheme operation

- Test materials will be distributed on a regular basis to the participants, who are required to return results by a given date.
- A value is assigned for each measurand, either before or after distribution; this value is not disclosed to participants until after the reporting deadline.
- The results are subjected to statistical analysis and/or converted into scores by the scheme provider, and participants are 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.

2.3.2 Structure of a single round

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

- the scheme provider organizes the preparation and validation of test material;
- the scheme provider distributes test samples on the regular schedule;
- the participants analyze the test materials and report results to the provider;
- the results are subjected to statistical analysis and/or scoring;
- the participants are notified of their performance;
- the scheme provider provides such advice as is available to poor performers, on request; and
- the scheme provider reviews the performance of the scheme during the particular round, and makes such adjustments as are necessary.

Note: Preparation for a round of the scheme will often have to be organized while the previous round is taking place.

2.4 Organization

- Day-to-day running of the scheme will be the responsibility of the scheme provider.
- The scheme provider must document all practices and procedures in their own quality manual, and a summary of relevant procedures must be supplied to all participants.
- The scheme provider should also keep the participants informed about the efficacy of the scheme as a whole, any changes that are being introduced, and how any problems have been dealt with.
- The operation of the scheme must be reviewed periodically (see below).
- Overall direction of the scheme must be overseen by an advisory committee having representatives (who should include practicing analytical chemists in the relevant field) from, for example, the scheme provider, contract laboratories (if any), appropriate professional bodies, participants, and end-users of analytical data. The advisory committee must also include a statistical expert.

2.5 Duties of the advisory committee

The advisory committee will consider and advise on the following subjects:

- the choice of the types of test materials, analytes, and the concentration ranges of the analytes
- the frequency of the rounds
- the scoring system and statistical procedures (including those used in homogeneity testing)
- the advice that can be offered to the participants
- specific and general problems arising during the operation of the scheme
- the instructions sent to participants
- the participants' format for reporting results
- the contents of the reports sent to participants
- other means of communicating with the participants
- comments from participants or end-users relating to the operation of the scheme
- the level of confidentiality appropriate to the scheme

2.6 Review of the scheme

- The operation of the scheme shall be reviewed regularly.
- The scheme provider shall review the outcomes of every round of the scheme, noting, for example, any strengths, weaknesses, specific problems, and opportunities for improvement.
- The provider and the advisory committee shall consider every aspect of the operation of the scheme, including the issues identified by the scheme provider's review of each round, usually at intervals of one year.
- A summary of this review shall be made available to participants and others as appropriate and agreed by the advisory committee.

2.7 Test materials

- The scheme provider shall arrange for the preparation of test materials. Preparation of test materials and other aspects of the scheme may be subcontracted, but the provider remains responsible and must exercise adequate control.
- The organization preparing the test material should have demonstrable experience in the area of analysis being tested.
- The test materials to be distributed in the scheme must be generally similar in type to the materials that are routinely analyzed (in respect of composition of the matrix and the concentration range, quantity, or level of the analyte).
- The bulk material prepared for the proficiency test must be sufficiently homogeneous and stable, in respect of each analyte, to ensure that all laboratories receive distribution units that do not differ to any consequential degree in mean analyte concentration (see Section 3.11). The scheme provider must clearly state the procedure used to establish the homogeneity of the test material.

  Note: While between-unit homogeneity is required to be sufficient, the participant should not assume that the distribution unit itself is sufficiently homogeneous for their particular analytical procedure. It is the responsibility of the participants to ensure that the test portion used for analysis is representative of the whole of the test material in the distribution unit.

- The quantity of material in a distribution unit must be sufficient for the analysis required, including any reanalysis where permitted by the scheme protocol.
- When unstable analytes are to be assessed, it may be necessary for the scheme provider to prescribe a date by, or on, which the analysis must be accomplished.

- Scheme providers must 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 any potential hazard involved.

  Note: "Appropriate action" includes, but is not limited to, compliance with specific legislation. Many countries also impose additional "duty of care", which may extend beyond legislative minimum requirements.

- The participants must be given, at the same time as the test materials, enough information about the materials, and any fitness-for-purpose criteria that will be applied, to allow them to select appropriate methods of analysis. This information must not include the assigned value.

2.8 Frequency of distribution

The appropriate frequency for the distribution of test materials shall be decided by the scheme provider with advice from the advisory committee (see Section 3.10). It will normally be between 2 and 10 rounds per year.

2.9 Assigned value

An assigned value is an estimate of the value of the measurand that is used for the purpose of calculating scores.

- An assigned value shall be determined by one of the following methods:
  - measurement by a reference laboratory*
  - the certified value(s) for a CRM used as a test material
  - direct comparison of the proficiency testing test material with CRMs
  - consensus of expert laboratories
  - formulation (i.e., value assignment on the basis of proportions used in a solution or other mixture of ingredients with known analyte content)
  - a consensus value (that is, a value derived directly from reported results)

  The assigned value will not be disclosed to the participants until after the reporting deadline for the results.

- The scheme provider must report the assigned value and an estimate of its uncertainty to the participants when reporting results and scores and must give sufficient details of how the assigned value and uncertainty were determined. Methods for determining the assigned value are discussed below (see Section 3.2).
- In sectors where empirical methods of analysis are used, the assigned value should normally be calculated from results obtained by using a clearly defined analytical procedure. Alternatively, the assigned value can be calculated from the results from two or more empirical methods shown to be effectively equivalent.
- Occasionally, it may be necessary for the scheme to use different assigned values for different methods, but this device should only be used to fulfil a clear necessity.
- Where an assigned value relies on an empirical method, participants must be told in advance which empirical procedure will be used for determining the assigned value.

*A "reference laboratory" in this context is a laboratory agreed by the scheme provider and advisory committee as providing reference values of sufficient reliability for the purpose of the scheme.

2.10 Choice of analytical method by participant

- Participants shall normally use the analytical method of their choice. In some instances, however, for example, where legislation so requires, participants may be instructed to use a specific documented method.
- Methods must be those used by the participant for routine work in the appropriate sector, and not versions of the method specially adapted for the proficiency test.

2.11 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 3.1.

> *Note*: The *z*-score based on a fitness-for-purpose criterion is the only type of score recommended in this protocol.

2.12 Performance criteria

For each analyte in a round, a criterion of performance must be set, against which the performance obtained by a laboratory can be judged. 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 usually be set to represent the best performance that typical methods are capable of providing (see Section 3.5).

2.13 Reporting of results by participants

- Participants must report results by the method and in the format required by the scheme.
- The scheme provider shall set a date by which results must be reported. Results submitted after the deadline must be rejected.
- Submitted results cannot be corrected or withdrawn.

> *Note*: The reason for this strict approach is that proficiency testing is meant to test every aspect of obtaining and producing an analytical result, including calculating, checking, and reporting a result.

2.14 Reports provided by scheme provider

- The scheme provider shall provide a performance report to each participant for each round.
- Reports issued to participants shall be clear and comprehensive and show the distribution of results from all laboratories together with participant's performance score.
- The test results as used by the scheme provider should also be available, to enable participants to check that their data have been correctly entered.
- Reports shall 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.
- Participants shall receive at least: (a) reports in clear and simple format, and (b) results of all laboratories in graphical form (e.g., as a histogram, bar chart, or other distribution plot) with appropriate summary statistics.

> *Note*: 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 hundreds of participants, each determining 20 analytes in any one round).

2.15 Liaison with participants

- On joining the scheme, participants shall be provided with detailed information, which shall describe
  - the range of tests available and the tests the participant has elected to undertake;
  - the method of setting performance criteria;
  - performance criteria applicable at the time of joining, unless criteria are set separately for each test material;
  - the method(s) of determining assigned values, including measurement methods where relevant;
  - a summary of the statistical procedures used to obtain participant scores;
  - information on interpreting scores;
  - conditions pertaining to participation (e.g., timeliness of reporting, avoidance of collusion with other participants, etc.);
  - the composition and method of selection of the advisory committee; and
  - contact details for the provider and any other relevant organization.

  Note: Communication with participants may be through any appropriate media, including, for example, periodic newsletters, the regular scheme review report, periodic open meetings, or electronic communication.

- Participants must be advised of any forthcoming changes in scheme design or operation.
- Advice must be available to poor performers, although this may be in the form of a list of consultants who are expert in the field.
- Participants who consider that their performance assessment is in error must be able to refer the matter to the scheme provider.
- There must be a mechanism whereby participants are able to comment on aspects of scheme operation and on problems with individual test materials so that participants contribute to the development of the scheme and to allow participants to alert the scheme provider to any unanticipated difficulty with test materials.

  Note: Feedback *from* participants should be encouraged.

2.16 Collusion and falsification of results

- It is the responsibility of the participating laboratories to avoid collusion or falsification of results. This shall be a documented condition of participation in a scheme, included in instructions to scheme participants.
- The scheme provider shall take due care to discourage collusion through appropriate scheme design. (For example, it could be advertised that more than one test material may occasionally be distributed within one round so that laboratories cannot compare results directly, and there should be no identifiable reuse of materials in successive rounds.)

  Note: Collusion, either between participants or between individual participants and the scheme provider, is contrary to professional scientific conduct and serves only to nullify the benefits of proficiency testing to customers, accreditation bodies, and analysts alike. Collusion is, therefore, to be strongly discouraged.

2.17 Repeatability

Reporting the mean of replicate determinations on proficiency test samples should be carried out only if this is the norm for routine work. (Procedures used by laboratories participating in proficiency testing schemes should simulate those used in routine sample analysis.)

> *Note*: Separate reporting of results replicated within laboratories is allowed as a possibility in proficiency tests, but is not recommended. If the practice is followed, scheme providers and participants must beware of misinterpreting repeatability standard deviations averaged over many participants. For example, the within-group sum of squares obtained by analysis of variance cannot be interpreted as an "average" repeatability variance when different analytical methods are in use.

2.18 Confidentiality

The degree of confidentiality to be exercised by the scheme provider and the participants with respect to scheme information and participant data shall be set out in the conditions of participation and notified to participants prior to joining the scheme.

> *Note*: In setting out the confidentiality conditions, organizers should consider the general benefit of open availability of general performance data for the analytical community, and are encouraged to provide for open publication of such information subject to due protection of individual participant information.

Unless stated otherwise in the conditions of participation:

- The scheme provider shall not disclose the identity of a participant to any third party, including other participants, without the express permission of the participant concerned.
- Participants shall be identified in reports by code only.

> *Note:* Random assignment of laboratory codes for each round prevents identification on the basis of history of participation, and is recommended where practicable.

- Participants may communicate their own results, including the regular scheme reports, privately to a laboratory accreditation or other assessment body when required for the purpose of assessment, or to clients (including the laboratory's parent organization, if applicable) for the purpose of demonstrating analytical capability.
- Participants may publish information on their own performance, but shall not publish comparative information on other participants, including score ranking.

PART 3: PRACTICAL IMPLEMENTATION

3.1 Conversion of participants' results into scores

3.1.1 The ethos of scoring

The 1993 Harmonized Protocol recommended the conversion of participants' results into *z*-scores, and experience in the intervening years has demonstrated the wide applicability and acceptance of the *z*-score in proficiency testing. A participant's result x is converted into a *z*-score according to the equation

$$z = (x - x_\mathrm{a})/\sigma_\mathrm{p} \qquad (1)$$

where x_a is the "assigned value", the scheme provider's best estimate of the value of the measurand (the true value of the concentration of the analyte in the proficiency testing material) and σ_p is the fitness-for-purpose-based "standard deviation for proficiency assessment". Guidance on the evaluation of x_a and σ_p are given below (see Sections 3.2–3.5).

Note 1: σ_p was designated the "target value" in the 1993 Harmonized Protocol [1]. This usage is now thought to be misleading.

Note 2: In ISO Guide 43 [4] and the Statistical Guide ISO 13528 [5], the symbol $\hat{\sigma}$ is used for standard deviation for proficiency assessment. σ_p is used here to underline the importance of assigning a range appropriate to a particular purpose.

The primary idea of the *z*-score is to make all proficiency test scores comparable, so that the significance of a score is immediately apparent, no matter what the concentration or identity of the analyte, the nature of the test material, the physical principle underlying the analytical measurement, or the organization providing the scheme. Ideally, a score of say $z = -3.5$, regardless of its origin, should have the same immediate implications for anybody, provider, participant, or end-user, involved in proficiency testing. This requirement is closely connected to the idea of fitness-for-purpose. In the equation defining *z*, the term $(x - x_\mathrm{a})$ is the error in the measurement. The parameter σ_p describes the standard uncertainty that is most appropriate for the application area of the results of the analysis, in other words, "fitness-for-purpose". That is not necessarily close to the uncertainty associated with the reported results. So although we can interpret *z*-scores on the basis of the standard normal distribution, we do not expect them to conform to that distribution.

The uncertainty that is fit for purpose in a measurement result depends on the application. For example, while a relative standard uncertainty [i.e., $u(x)/x$] of 10 % is probably adequate for many environmental measurements, a much smaller relative uncertainty is required for assaying a shipment of scrap containing gold to determine its commercial value. But there is more to it than that. Deciding on a fit-for-purpose uncertainty is a trade-off between costs of analysis and costs of making incorrect decisions. Obtaining smaller uncertainty requires disproportionately larger expenditure on analysis. But employing methods with greater uncertainty means a greater likelihood of making an expensive incorrect decision based on the data. Fitness-for-purpose is defined by the uncertainty that balances these factors, i.e., that minimizes the expected total loss [10]. Analysts and their customers do not usually make a formal mathematical analysis of the situation, but should at least agree as to what comprises fitness-for-purpose for each specific application.

3.1.2 How should z-scores be interpreted?

It is important to emphasize that the interpretation of *z*-scores is *not* generally based on summary statistics that describe the observed participant results. Instead, it uses an assumed model based on the scheme provider's fitness-for-purpose criterion, which is represented by the standard deviation for proficiency assessment σ_p. Specifically, interpretation is based on the normal distribution $x \sim N(x_\mathrm{true}, \sigma_\mathrm{p}^2)$, where x_true is the true value for the quantity being measured. Under this model regime, and assuming that the assigned value is very close to x_true so that the *z*-scores follow the standard normal distribution:

- A score of zero implies a perfect result. This will happen rarely even in the most competent laboratories.
- Approximately 95 % of *z*-scores will fall between –2 and +2. The sign (i.e., – or +) of the score indicates a negative or positive error respectively. Scores in this range are commonly designated "acceptable" or "satisfactory".
- A score outside the range from –3 to 3 would be very unusual and is taken to indicate that the cause of the event should be investigated and remedied. Scores in this class are commonly designated "unacceptable" or "unsatisfactory", although a nonpejorative phrase such as "requiring action" is preferable.
- Scores in the ranges –2 to –3 and 2 to 3 would be expected about 1 time in 20, so an isolated event of this kind is not of great moment. Scores in this class are sometimes designated "questionable".

Few if any laboratories conform exactly with the above. Most participants will operate with a biased mean and a run-to-run standard deviation that differs from σ_p. Some will generate extreme results due to gross error. However, the model serves as a suitable guide to action on the z-scores received by *all* participants, for the following reasons. A biased mean or a standard deviation greater than σ_p will, in the long run, always produce a greater proportion of results giving rise to $|z| > 2$ and $|z| > 3$ than the standard normal model (that is, about 0.05 and 0.003, respectively). This will correctly alert the participant to the problem. Conversely, a participant with an unbiased mean and standard deviation equal to or smaller than σ_p will produce a small proportion of such results, and will correctly receive few adverse reports.

3.2 Methods of determining the assigned value

There are several possible approaches that the proficiency testing provider can employ to determine the assigned value and its uncertainty. All have strengths and weaknesses. Selection of the appropriate method for assignment in different schemes, and even rounds within a scheme or series, will therefore depend on the purposes of the scheme. In selecting methods of value assignment, scheme organizers and advisory committees should consider the following.

- the costs to organizer and participants—high costs may deter participation and thereby reduce the effectiveness of the scheme
- any legal requirements for consistency with reference laboratories or other organizations
- the need for independent assigned values to provide a check on bias for the population as a whole
- any specific requirement for traceability to particular reference values

Note: Implementation of metrological traceability by participants is expected as an essential element of good analytical QA. Where metrological traceability and appropriate QA/QC methods—particularly validation using appropriate matrix CRMs—are adequately implemented, good consensus and low dispersion of results are the expected outcome. Simply observing the dispersion of results (and in particular, the fraction of laboratories achieving acceptable scores) is accordingly a direct test of effective traceability, irrespective of the assigned value. However, testing against an independently assigned traceable value can provide a useful additional check on effective traceability.

3.2.1 Measurement by a reference laboratory

In principle, an assigned value and uncertainty may be obtained by a suitably qualified measurement laboratory using a method with sufficiently small uncertainty. For most practical purposes, this is exactly equivalent to use of a CRM (below). It is advantageous in that the material is effectively tailored to the scheme requirements. The principal disadvantage is that it may require disproportionate effort and cost if, for example, substantial investigations are required to validate the methodology for the material in question or to eliminate the possibility of significant interferences.

3.2.2 Use of a certified reference material

If a CRM is available in sufficient amounts for use in a proficiency test, the certified value(s) and associated uncertainty can be used directly. This is quick and simple to implement, and (usually) provides a value independent of the participant results. Appropriate traceability for the reference value is also automatically provided (by definition). There are, however, disadvantages. Natural matrix CRMs are not usually available in sufficient amounts and/or at suitable cost to use regularly in proficiency testing schemes. They may be easily recognizable by the participants, who would then be able to infer the certified value. Finally, although proficiency tests are generally valuable, they are most valuable in analytical sectors where reference materials are scarce or not available.

3.2.3 Direct comparison of the proficiency testing material with certified reference materials

In this method, the test material is analyzed several times alongside appropriate CRMs in a randomized order under repeatability conditions (i.e., in a single run) by a method with suitably small uncertainty. Provided that the CRMs are closely comparable with the prospective proficiency testing material in respect of the matrix and the concentration, speciation, and compartmentation of the analyte, the result for the proficiency testing material, determined via a calibration function based on the certified values of the CRMs, will be traceable to the CRM values and through them to higher standards. The uncertainty will incorporate only terms due to the uncertainties of the CRMs and repeatability error of the analysis.

Note: This practice is described in ISO 13528 [5]. It is identical to performing a measurement using matched CRMs as calibrants, and might, therefore, reasonably be described as "measurement using matched calibration materials".

In practice, it is difficult to determine whether the CRMs are sufficiently similar in all respects to the proficiency testing material. If they are dissimilar, an extra contribution must be included in the uncertainty calculation for the assigned value. It is difficult to determine the magnitude of this extra contribution. As before, proficiency tests are most valuable in analytical sectors where reference materials are not available.

3.2.4 Consensus of expert laboratories

The assigned value is taken as the consensus of a group of expert laboratories that achieve agreement on the proficiency testing material by the careful execution of recognized reference methods. This method is particularly valuable where operationally defined ("empirical") parameters are measured, or, for example, where routine laboratory results are expected to be consistent with results from a smaller population of laboratories identified by law for arbitration or regulation. It also has the advantage of providing for cross-checking among the expert laboratories, which helps to prevent gross error.

In practice, however, the effort required to achieve consensus and a usefully small uncertainty is about the same as that required to certify a reference material. If the reference laboratories used a routine procedure to analyze the proficiency testing material, their results would tend to be no better on average than those of the majority of participants in the proficiency testing proper. Further, as the number of available reference laboratories is perforce small, the uncertainty and/or variability of a subpopulation's consensus might be sufficiently large to prejudice the proficiency test.

Where a consensus of expert laboratories is used, the assigned value and associated uncertainty are assessed using an appropriate estimate of central tendency (usually, the mean or a robust estimate thereof). The uncertainty of the assigned value is then based either on the combined reported uncertainties (if consistent) or on the appropriate statistical uncertainty combined with any additional terms required to account for calibration chain uncertainties, matrix effects, and any other effects.

3.2.5 Formulation

Formulation comprises the addition of a known amount or concentration of analyte (or a material containing the analyte) to a base material containing none. The following circumstances have to be considered.

- The base material must be effectively free of the analyte, or its concentration must be accurately known.
- It may be difficult to obtain sufficient homogeneity (see Section 3.11) when a trace analyte is added to a solid base material.
- Even when the speciation is appropriate, the added analyte may be more loosely bonded to the matrix than the analyte native in typical test materials, and hence the recovery of the added analyte may be unrealistically high.

Providing that these problems can be overcome, the assigned value is determined simply from the proportions of the materials used and the known concentrations (or purity if a pure analyte is added). Its uncertainty is normally estimated from the uncertainties in purity or analyte concentrations of the materials used and gravimetric and volumetric uncertainties, though issues such as moisture content and other changes during mixing must also be taken into account if significant. The method is relatively easy to execute when the proficiency testing material is a homogeneous liquid and the analyte is in true solution. However, it may be unsuitable for solid natural materials where the analyte is already present ("native" or "incurred").

3.2.6 Consensus of participants

The consensus of the participants is currently the most widely used method for determining the assigned value: Indeed, there is seldom a cost-effective alternative. The idea of consensus is not that all of the participants agree within bounds determined by the repeatability precision, but that the results produced by the majority are unbiased and their dispersion has a readily identifiable mode. To derive a most probable value for the measurand (i.e., the assigned value) we use an appropriate measure of the central tendency of the results and we (usually) use its standard error as the estimate of its uncertainty (see Section 3.3).

The advantages of participant consensus include low cost, because the assigned value does not require additional analytical work. Peer acceptance is often good among participants because no one member or group is accorded higher status. Calculation of the value is usually straightforward. Finally, long experience has shown that consensus values are usually very close, in practice, to reliable reference values provided by formulation, expert laboratory consensus, and reference values (whether from CRMs or reference laboratories).

The principal disadvantages of participant consensus values are, first, that they are not independent of the participant results and second, that their uncertainty may be too large where the number of laboratories is small. The lack of independence has two potential effects: (i) bias for the population as a whole may not be detected promptly, as the assigned value will follow the population; (ii) if the majority of results are biased, participants whose results are unbiased may unfairly receive extreme *z*-scores. In practice, the former is rare except in small populations using the same method; the existence of several distinct subpopulations is a more common problem. Both providers of proficiency tests and participants must accordingly be alert to these possibilities (though they should be equally alert to the possibility of error in any other value assignment method). The situation is usually quickly rectified once it is recognized. It is one of the benefits of proficiency testing that participants can be made aware of unrecognized *general* problems as well as those involving particular laboratories.

The limitations induced by small group sizes are often more serious. When the number of participants is smaller than about 15, even the statistical uncertainty on the consensus (identified as the standard error) will be undesirably high, and the information content of the *z*-scores will be correspondingly reduced.

Despite the apparent disadvantages, however, there is a large body of experience demonstrating that proficiency tests operate very well by using the consensus, so long as organizers are alive to the possibility of occasional difficulties and apply appropriate methods of calculation. Exact methods of estimating a consensus from participants' results are accordingly discussed in detail below.

3.3 Estimating the assigned value as the consensus of participants' results

3.3.1 Estimates of central tendency

If the results of the participants in a round are unimodal and, outliers aside, reasonably close to symmetric, the various measures of central tendency are nearly coincident. Accordingly, we feel confident about taking one of them, such as the mode, the median, or a robust mean, as the assigned value. We need to use an estimation method that is insensitive to the presence of outliers and heavy tails to avoid undue influence from poor results, and this is why the median or a robust mean is valuable.

Robust statistics are based on the assumption that the data are a sample from an essentially normal distribution contaminated with heavy tails and a small proportion of outliers. The statistics are calculated by downweighting the data points that are distant from the mean and then compensating for the downweighting. There are many versions of robust statistics [5,11]. The median is a simple type of robust mean. The Huber robust mean, obtained by the algorithm recommended by the Analytical Methods Committee (AMC) [11], and by ISO 5725 and ISO 13528 as "algorithm A", makes more use of the information in the data than the median does, and, consequently, in most circumstances has a somewhat smaller standard error. The median, however, is more robust when the frequency distribution is strongly skewed. The robust mean is, therefore, preferred when the distribution is close to symmetric. The mode is not defined exactly for samples from continuous distributions, and special methods are required to estimate it [12]. Nonetheless, the mode may be especially useful when bimodal or multimodal results are obtained (see Appendix 3).

A recommended scheme for estimating the consensus and its uncertainty is outlined below. An element of judgement, based in expertise in analytical chemistry and statistics, is written into this scheme; that is a strength rather than a weakness and is regarded as essential. This is because it is difficult or impossible to devise a set of rules that can be executed mechanically to provide an appropriate consensus for any arbitrary data set.

3.3.2 Recommended scheme for obtaining a consensus value and its uncertainty

The recommended scheme for obtaining an assigned value x_a and its uncertainty by consensus is given in the procedure set out in the frame below. The rationale for certain details is discussed in Section 3.3.3. Examples of the use of this scheme are given in Appendix I.

Recommendation 1

a. Exclude from the data any results that are identifiably invalid (e.g., if they are expressed in the wrong units or obtained by using a proscribed method) or are extreme outliers (for example, outside the range of ±50 % of the median).

b. Examine a visual presentation of the remaining results, by means of a dot plot [for small (n < 50) data sets], bar chart, or histogram (for larger data sets). If outliers cause the presentation of the bulk of the results to be unduly compressed, make a new plot with the outliers deleted. If the distribution is, outliers aside, apparently unimodal and roughly symmetric, go to (c), otherwise go to (d).

c. Calculate the robust mean $\hat{\mu}_{\text{rob}}$ and standard deviation $\hat{\sigma}_{\text{rob}}$ of the n results. If $\hat{\sigma}_{\text{rob}}$ is less than about $1.2\sigma_p$, then use $\hat{\mu}_{\text{rob}}$ as the assigned value x_a and $\hat{\sigma}_{\text{rob}}/\sqrt{n}$ as its standard uncertainty. If $\hat{\sigma}_{\text{rob}} > 1.2\sigma_p$, go to (d).

d. Make a kernel density estimate of the distribution of the results using normal kernels with a bandwidth h of $0.75\sigma_p$. If this results in a unimodal and roughly symmetric kernel density, and the mode and median are nearly coincident, then use $\hat{\mu}_{\text{rob}}$ as the assigned value x_a and $\hat{\sigma}_{\text{rob}}/\sqrt{n}$ as its standard uncertainty. Otherwise, go to (e).

e. If the minor modes can safely be attributed to outlying results, and are contributing less than about 5 % to the total area, then still use $\hat{\mu}_{\text{rob}}$ as the assigned value x_a and $\hat{\sigma}_{\text{rob}}/\sqrt{n}$ as its standard uncertainty. Otherwise, go to (f).

f. If the minor modes make a considerable contribution to the area of the kernel, consider the possibility that two or more discrepant populations are represented in the participants' results. If it is possible to infer from independent information (e.g., details of the participants' analytical methods) that one of these modes is correct and the others incorrect, use the selected mode as the assigned value x_a and its standard error as its standard uncertainty. Otherwise, go to (g).

g. The methods above having failed, abandon the attempt to determine a consensus value and report no individual laboratory performance scores for the round. It may still be useful, however, to provide the participants with summary statistics on the data set as a whole.

3.3.3 Notes on the rationale of the scheme for determining the assigned value

The rationale for the above scheme is as follows:

The use of $\hat{\sigma}_\mathrm{rob}/\sqrt{n}$ as the standard uncertainty of the assigned value is open to objection on theoretical grounds, because the influence of some of the n results is downweighted in calculating $\hat{\sigma}_\mathrm{rob}$ and its sampling distribution is complex. It is, however, one of the methods recommended in ISO 13528. In practice, $u(x_\mathrm{a}) = \hat{\sigma}_\mathrm{rob}/\sqrt{n}$ is only used as a rough guideline to the suitability of the assigned value, and the theoretical objection is of little concern.

In (b) above, we expect $\hat{\sigma}_\mathrm{rob} \approx \sigma_\mathrm{p}$ as the participants will be attempting to achieve fitness-for-purpose. If we find that $\hat{\sigma}_\mathrm{rob} > 1.2\sigma_\mathrm{p}$, it is a reasonable assumption either that laboratories are having difficulty achieving the required reproducibility precision in results from a single population, or that two or more discrepant populations may be represented in the results. A kernel density may help to decide between these possibilities. Whether or not the latter situation results in two (or more) modes depends on the separation of the means of the populations and the number of results in each sample.

Using a bandwidth h of $0.75\sigma_\mathrm{p}$ to construct kernel densities is a compromise that inhibits the incidence of artefactual modes without unduly increasing the variance of the kernel density in relation to $\hat{\sigma}_\mathrm{rob}$.

3.4 Uncertainty on the assigned value

If there is an uncertainty $u(x_\mathrm{a})$ in the assigned value x_a, and a participant is performing according to the standard deviation σ_p defining fitness-for-purpose, the uncertainty on a participant's deviation from x_a would be $\sqrt{u^2(x_\mathrm{a}) + \sigma_\mathrm{p}^2}$, so we might expect to see z-scores with a dispersion other than $N(0, 1)$. It is, therefore, appropriate to compare $u^2(x_\mathrm{a})$ with σ_p^2 to check that the former is not having an adverse effect on z-scores. For instance, if $u^2(x_\mathrm{a}) = \sigma_\mathrm{p}^2$, the z-scores would be dilated by a factor of about 1.4, which would be an unacceptable outcome. On the other hand, if $u^2(x_\mathrm{a}) > \sigma_\mathrm{p}^2$, the dilation factor would be about 1.05, the effect of which would be negligible for practical purposes. Accordingly, it is recommended that z-scores are not presented to participants in an unqualified form if it is found that $u^2(x_\mathrm{a}) > 0.1\sigma_\mathrm{p}^2$. (The factor of 0.1 is of appropriate magnitude, but its exact value is essentially arbitrary and should be considered by the scheme provider.) If the inequality were exceeded somewhat (but not greatly), the scheme could issue the z-scores with a warning qualification attached to them, for example, by labeling them "provisional" with a suitable explanation. Therefore, proficiency testing providers would need to nominate a suitable value of l in the expression $u^2(x_\mathrm{a}) = l\ \sigma_\mathrm{p}^2$, higher than which no z-scores would be calculated.

ISO 13528 [5] refers to a modified z-score z' given by $z' = \dfrac{x - x_\mathrm{a}}{\sqrt{u^2(x_\mathrm{a}) + \sigma_\mathrm{p}^2}}$ that could be used when the uncertainty of the assigned value was non-negligible. However, z' is *not* recommended for use in this protocol. While it would tend to give values similar to proper z-scores in dispersion, the use of z' would disguise the fact that the uncertainty on the assigned value was unsuitably high. The current recommendation is therefore as follows.

Recommendation 2

The proficiency testing provider should nominate a multiplier $0.1 < l < 0.5$ appropriate for the scheme and, having evaluated $u^2(x_a) + \sigma_p^2$ for a round, act as follows:

- if $u^2(x_a) + \sigma_p^2 \leq 0.1$, issue unqualified z-scores;
- if $0.1 < u^2(x_a) + \sigma_p^2 \leq l$, issue qualified z-scores (such as "provisional z-scores");
- if $u^2(x_a) + \sigma_p^2 > l$, do not issue z-scores.

Note: In the inequality $0.1 < l < 0.5$, the limits can be modified somewhat to meet the exact requirements of particular schemes.

3.5 Determination of the standard deviation for proficiency assessment

The standard deviation for proficiency assessment σ_p is a parameter that is used to provide a scaling for the laboratory deviations $(x - x_a)$ from the assigned value and thereby define a z-score (Section 3.1). There are several ways in which the value of the parameter can be determined, and their relative merits are discussed below.

3.5.1 Value determined by fitness-for-purpose

In this method, the proficiency testing provider determines a level of uncertainty that is broadly accepted as appropriate by the participants and the end-users of the data for the sector of application of the results, and defines it in terms of σ_p. By "appropriate", we mean that the uncertainty is small enough that decisions based on the data will only rarely be incorrect, but not so small that the costs of analysis will be unduly high. A suggested definition of this "fitness-for-purpose" is that it should comprise the uncertainty that minimizes the combined costs of analysis and the financial penalties associated with incorrect decisions multiplied by their probabilities of occurrence [10]. It must be emphasized that σ_p does not here represent a general idea of how laboratories are performing, but how they ought to be performing to fulfil their commitment to their clients. The numerical value of the parameter should be such that the resultant z-scores can be interpreted by reference to the standard normal distribution. It will probably be determined by professional judgement exercised by the advisory committee of the scheme. In some analytical sectors, there is already an acknowledged working standard for fitness-for-purpose. For instance, in the food sector, the Horwitz function is often regarded as defining fitness-for-purpose as well as being simply descriptive [13].

However the value of the parameter is arrived at, it will have to be determined and publicized in advance of the distribution of the proficiency testing materials so that participants can check whether their analytical procedures conform with it. In some schemes, the possible range of the analyte concentration is small and a single level of uncertainty can be specified to cover all eventualities. A complication arises in instances where the concentration of the analyte can vary over a wide range. As the assigned value is not known in advance by the participants, the fitness-for-purpose criterion has to be specified as a function of concentration. The most common approaches are as follows:

- Specify the criterion as a relative standard deviation (RSD). Specific σ_p values are then obtained by multiplying this RSD by the assigned value.
- Where there is a lower limit of interest in the analytical result, set an RSD applicable over a specified range in conjunction with a (lower) limiting value for σ_p. For example, in the determination of the concentration of lead in wine, it would be prudent to aim for an RSD of 20 % over a wide range of analyte concentrations, but at concentrations well below the maximum allowable concentration x_{max} such a level of precision would be neither necessary nor cost-effective. That fact could be recognized by formulating the fitness-for-purpose criterion as a function in the form

$$\sigma_{\mathrm{p}} = x_{\mathrm{max}}/m + 0.2x_{\mathrm{a}} \tag{2}$$

where *f* is a suitable constant. If *f* were set at 4, for example, σ_{p} would never be lower than $x_{\mathrm{max}}/4$.

- Specify a general expression of fitness-for-purpose, such as the Horwitz function [13], namely, (using current notation):

$$\sigma_{\mathrm{p}} = 0.02x_{\mathrm{a}}^{0.8495} \tag{3}$$

where x_{a} and σ_{p} are expressed as mass fraction. Note that the original Horwitz relationship loses applicability at concentrations lower than about 10 ppb (ppb = 10^{9} mass fraction), and a modified form of the function has been recommended [14].

3.5.2 Legally defined value

In some instances, a maximum reproducibility standard deviation for analytical results for a specific purpose is set by legislation or international agreement. This value may be usable as a value for σ_{p}. Similarly, if a limit of permitted error has been set, this may be used to set σ_{p} by, for example, dividing by the appropriate value of student's *t* if a level of confidence is also available. However, it may still be preferable to use a value of σ_{p} lower than the legal limit. That is a matter for the provider and advisory committee of the proficiency testing scheme.

3.5.3 Other approaches

Scoring in some proficiency testing schemes is not based on the idea of fitness-for-purpose, which greatly diminishes the value of scoring. While such scoring methods are covered by ISO Guide 43 [4] (and discussed in the previous version of this Harmonized Protocol [1]), they are not recommended here for chemical proficiency testing. There are essentially two versions of such scoring systems. In one of these, the value of σ_{p} is determined by expert perception of laboratory performance for the type of analysis under consideration. Clearly, how laboratories perform could be better or worse than fit-for-purpose, so the scoring system tells us only which laboratories are out of line with other participants, not whether any of them are good enough. Another version of this method, seemingly more authoritative because it relies on standard statistical ideas, is to use the robust standard deviation of the participants' results in a round as σ_{p}. The outcome of that strategy is that in every instance about 95 % of the participants receive an apparently acceptable *z*-score. That is a comforting outcome for both the participants and the scheme providers but, again, it serves only to identify results that are out of line. There is an added difficulty that the value used for σ_{p} will vary from round to round so there is no stable base for comparison of scores between rounds. Although the method can be improved by using a fixed value derived by combining results of several rounds, it still provides no incentive for laboratories producing results that are unfit for purpose to improve their performance.

There may be circumstances when it is justifiable for a proficiency testing scheme not to provide guidance on fitness-for-purpose. That is the case when the participants carry out their routine work for a variety of different purposes, so there can be no universally applicable fitness-for-purpose criterion. In such conditions, it would be better for the proficiency testing provider to provide no scoring at all, but just give an assigned value (with its uncertainty) and perhaps the laboratory error. (This is sometimes provided in terms of relative error, the so-called "Q score"). Any scoring that is provided should be clearly indicated as "for informal use only", to minimize the incidence of accreditation assessors or prospective clients making incorrect judgements based on the scores. Individual participants in such schemes then have to provide their own criteria of fitness-for-purpose, and a scheme for carrying that out is provided below (see Section 3.6 and Appendix 6).

Recommendation 3

Wherever possible, the proficiency testing scheme should use for σ_p, the standard deviation for proficiency assessment, a value that reflects fitness-for-purpose for the sector. If there is no single level that is generally appropriate, the provider should refrain from calculating scores, or should show clearly on the reports that the scores are for informal descriptive use only and not to be regarded as an index of performance of the participants.

3.5.4 Modified z-score for individual requirements

Some proficiency testing schemes do not operate on a "fitness-for-purpose" basis. The scheme provider calculates a score from the participants' results alone (i.e., with no external reference to actual requirements). Alternatively, a participant may find that the fitness-for-purpose criterion used by the scheme provider is inappropriate for certain classes of work undertaken by the laboratory. In fact, it would not be unusual for a laboratory to have a number of customers, wanting the same analyte determined in the same material, but each having a different uncertainty requirement.

In proficiency testing schemes operating on either of these bases, participants can calculate scores based on their own fitness-for-purpose requirements. That can be accomplished in a straightforward manner. The participant should agree on a specific fitness-for-purpose criterion σ_{ffp} with a customer for each specific application, and use that to calculate the corresponding modified z-score, given by

$$z_L = (x - x_a)/\sigma_{\mathrm{ffp}} \tag{4}$$

to replace the conventional z-score [15]. The criterion σ_{ffp} could be expressed as a function of concentration if necessary. It should be used like the sigma value in a z-score, that is, it should be in the form of a standard uncertainty that represents the agreed fitness-for-purpose. If there were several customers with different accuracy requirements, there could be several valid scores derived from any one result. The modified z-scores can be interpreted in exactly the manner recommended for z-scores (see Appendix 6).

3.6 Participant data reported with uncertainty

This Protocol does not recommend the reporting of participants' uncertainty of measurement with the result. This recommendation is consistent with ISO Guide 43. Indeed, relatively few proficiency testing schemes for analytical chemistry currently require participants' results to be accompanied by an uncertainty estimate. This is principally because it is assumed that, after careful expert consideration, schemes commonly set a σ_p value that represents fitness-for-purpose over a whole application sector. This optimal uncertainty requirement is, therefore, implicit in the scheme. Participants are expected to perform in a manner consistent with this specification and, therefore (in this context), do not need to report uncertainties explicitly. Those performing in accordance with the scheme's requirement will usually receive z-scores in the range ±2. Those participants with significantly underestimated uncertainties are far more likely to receive "unacceptable" z-scores. In other words, correctly estimated uncertainties would be expected mostly to be similar to the σ_p value, and underestimates would tend to result in poor z-scores. In such circumstances, uncertainty reporting does not add to the value of the scheme. Further, proficiency testing schemes to date have been extremely effective without uncertainty data from participants; it follows that uncertainty data from participants is not required for the improvement of routine analytical performance.

However, the circumstances outlined above may not be universally applicable. Laboratories working to their own fitness-for-purpose criteria should, therefore, be judged by individual criteria rather than the generic σ_p value for the scheme. Further, uncertainty data are increasingly required by customers of laboratories, and laboratories should accordingly be checking their procedures for doing so. The following sections accordingly discuss three important issues relating to use of participant uncertainties; the determination of consensus values, the use of scoring as a check on reported uncertainty, and the use of participant uncertainty in assessing individual fitness-for-purpose.

3.6.1 Consensus values

Where uncertainty estimates are available, scheme providers may need to consider how a consensus is best identified when the participants report data with uncertainties, and how the uncertainty of that consensus is best estimated. The naïve version of the problem is establishing a consensus and its uncertainty from a set of unbiased estimates of a measurand, each with a different uncertainty. The reality is that: (a) there are often discordant results among those reported (that is, the data comprise samples from distributions with different means); and (b) the uncertainty estimates are often incorrect and, in particular, those coupled with outlying results are likely to be much too small.

At present, there are no well-established methods for providing robust estimates of mean or dispersion for interlaboratory data with variable uncertainties. This area is, however, under active development, and several interesting proposals have been discussed [16,17]. For example, methods based on kernel density estimation [12] currently seem likely to be productive.

Fortunately, most participants in a given scheme are working to similar requirements and would be expected to provide similar uncertainties. Under these circumstances, weighted and unweighted estimates of central tendency are very similar. Robust unweighted estimates may, therefore, be applied, with the advantage of less sensitivity to substantial underestimates of uncertainty or distant outliers.

Given that these topics are still under active development, and that uncertainty estimates are somewhat unreliable, it is recommended that unweighted robust methods be used for assigned value calculation.

Recommendation 4

Even when uncertainty estimates are available, unweighted robust methods (i.e., methods taking no account of the individual uncertainties) should be used to obtain the consensus value and its uncertainty, according to the methods described in Sections 3.3 and 3.4.

3.6.2 The zeta score

ISO 13528 defines the zeta score (ζ) for scoring results and reported uncertainties as follows:

$$\zeta = (x - x_\mathrm{a}) \Big/ \sqrt{u^2(x) + u^2(x_\mathrm{a})} \tag{5}$$

where $u(x)$ is the reported standard uncertainty in the reported value x and $u(x_a)$ the standard uncertainty for the assigned value. The zeta score provides an indication of whether the participant's estimate of uncertainty is consistent with the observed deviation from the assigned value. The interpretation is similar to the interpretation of z-scores; absolute values over 3 should be regarded as cause for further investigation. The cause might be underestimation of the uncertainty $u(x)$, but might also be due to gross error causing the deviation x-x_a to be large. The latter condition would usually be expected to result in a high z-score, so it is important to consider z and zeta scores together. Note, too, that persistently low zeta scores over a period of time might indicate over-estimation of uncertainty.

Note: ISO 13528 defines additional scoring methods which use expanded uncertainty; reference to ISO 13528 is recommended if this is considered appropriate by the scheme advisory committee.

3.6.3 Scoring results that are reported with uncertainty

It is easy for a participant to use the zeta score to check their own estimates of uncertainty. At the present time, however, we are considering actions taken by organizers of proficiency testing schemes.

The question at issue is whether the scheme provider (rather than individual participants) should attempt to take uncertainty into account in converting the raw results into scores. There is no particular difficulty in doing this—it is merely a question of whether the outcome would be useful. The balance of benefits falls against the schemes calculating zeta scores. All schemes are encouraged to provide a diagram showing the participants' results and scores. Such a diagram based on zeta scores would be ambiguous because the results could not be usefully represented in a two-dimensional plot. A particular zeta score (say, –3.7) could have resulted either from a large error and a large uncertainty, or a small error and a proportionately small uncertainty. Moreover, the scheme organizer has no means of judging whether a participant's submitted uncertainty value is appropriate to their needs, so that the zeta scores so produced would be of unknown value in assessing the participants' results.

Recommendation 5

Schemes should not provide zeta scores unless there are special reasons for doing so. Where a participant has requirements inconsistent with that of the scheme, the participant may calculate zeta scores or the equivalent.

3.7 Scoring results near the detection limit

Many analytical tasks involve measuring concentrations of analyte that are close to the detection limit of the method, or even at exactly zero. Proficiency testing these methods should ideally mimic true life; the test materials should contain typically low concentrations of analyte. There are, however, difficulties in applying the usual z-scoring method to results of such tests. These difficulties are partly caused by the following data recording practices:

- Many practicing analysts finding a low result will record "default" results such as "not detected" or "less than c_L", where c_L is an arbitrarily determined limit. Such results, while possibly fit-for-purpose, cannot be converted into a z-score. z-Scores require the analytical result to be on an interval scale or ratio scale. Replacing the default result in an arbitrary way (e.g., by zero, or one-half of the detection limit) is not recommended.
- Some proficiency testing schemes avoid the difficulty by not processing default results. If many participants are working close to their detection limits, regardless of whether they provide a default result, it becomes difficult to estimate a valid consensus for the assigned value. The distribution of the apparently valid results tends to have a strong positive skew, and most kinds of average tend to have a high bias.

These difficulties can be circumvented by working at somewhat higher concentrations than typically found in the materials of interest. This practice is not entirely satisfactory, because the samples are then unrealistic. If participants recorded the actual result found, plus the (correct) uncertainty of the result, it would be possible in principle to estimate a valid consensus with an uncertainty. While this is to be recommended where possible, other factors such as established practice in customer reporting requirements make it an unlikely scenario for routine analysis. It therefore seems that z-scores can be used at low concentrations only when all the following apply:

- The participants record the actual results found.
- The assigned value is independent of the results. That might be achievable if the assigned value were known to be zero or very low, or could be determined by formulation or a reference laboratory (see Section 3.2).
- The standard deviation for proficiency assessment is an independent fitness-for-purpose criterion; its value would then be predetermined, that is, independent of the participants' results. This would be relatively straightforward to effect.

At present, there is no alternative well-established scoring system for low results in proficiency tests, and the subject is still very much under discussion. If the results are required to be essentially binomial ($\leq x$ or $>x$), then a scoring system could be devised based on the proportion of correct outcomes, but it is bound to be less powerful (i.e., less information-rich) than the z-scoring system. A mixed system (capable of dealing with a mixture of binomial, ordinal, and quantitative results) cannot readily be envisaged.

3.8 Caution in the uses of *z*-scores

Appropriate uses of z-scores by participants and end-users are discussed in some detail in Appendices 6 and 7. The following words of caution are addressed to providers.

It is common for several different analyses to be required within each round of a proficiency test. While each individual test furnishes useful information, it is tempting to determine a single figure of merit that will summarize the overall performance of the laboratory within a round. 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 provision of combination scores in reports to participants is not recommended, but it is recognized that such scores may have specific applications, if they are based on sound statistical principles and issued with proper cautionary notice. The procedures that may be used are described in Appendix 4.

It is especially emphasized 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.

3.9 Classification, ranking, and other assessments of proficiency data

Classification of laboratories is not the aim of proficiency testing, and is best avoided by providers as it is more likely to cause confusion than illumination. The replacement of a continuous measure such as a z-score by a few named classes has little to commend it from the scientific point of view: Information is thrown away. Consequently, classification is not recommended in proficiency tests. Decision limits based on z-scores may be used as guidelines where necessary. For example, a z-score outside the range ±3 could be regarded as requiring investigation leading, where necessary, to modification of procedures. Even so, such limits are arbitrary. A score of 2.9 should be almost as worrying as a score of 3.1. Moreover, these are matters for individual participants rather than scheme providers.

Ranking laboratories on their absolute z-scores obtained in a round of a scheme, to form a league table, is even more invidious than classification. A participant's rank is much more variable from round to round than the scores they are derived from, and the laboratory with the smallest absolute score in a round is unlikely to be "the best".

Recommendation 6

Proficiency scheme providers, participants, and end-users should avoid classification and ranking of laboratories on the basis of their *z*-scores.

3.10 Frequency of rounds

The appropriate distribution frequency is a balance between a number of factors of which the most important are

- the difficulty of executing effective analytical QC;
- the laboratory throughput of test samples;
- the consistency of the results in the particular field of work covered by the scheme;
- the cost/benefit of the scheme;
- the availability of CRMs in the analytical sector; and
- the rate of change of analytical requirements, methodology, instrumentation, and staff in the sector of interest.

Objective evidence about the influence of round frequency on the efficacy of proficiency testing is very sparse. Only one reliable study on frequency has been reported [18], and that showed (in a particular scheme) that changing the round frequency from three to six per year had no significant effect (beneficial or otherwise) on the participants' performance.

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-around time of test samples and results. It might also encourage the belief that the proficiency testing scheme can be used as a substitute for IQC, an idea that is strongly to be discouraged. If the period between distributions extends much beyond four months, there will be unacceptable delays in identifying and correcting analytical problems, and the impact of the scheme on the participants could be small. There is little practical value, in routine analytical work, in proficiency tests undertaken much less than twice a year.

3.11 Testing for sufficient homogeneity and stability

3.11.1 Testing for "sufficient homogeneity"

Materials prepared for proficiency tests and other interlaboratory studies are usually heterogeneous to some degree, despite best efforts to ensure homogeneity. When such a bulk material is split for distribution to various laboratories, the units produced vary slightly in composition among themselves. This protocol requires that this variation is sufficiently small for the purpose. A recommended procedure is provided in Appendix 1. The rationale for this procedure is discussed in the following paragraphs.

When we test for so-called "sufficient homogeneity" in such materials, we are seeking to show that this variation in composition among the distributed units (characterized by the sampling standard deviation σ_{sam}) is negligible in relation to variation introduced by the measurements conducted by the participants in the proficiency test. As we expect the standard deviation of interlaboratory variation in proficiency tests to be approximated to by σ_{p}, the "standard deviation for proficiency assessment", it is natural to use that criterion as a reference value. The 1993 Harmonized Protocol [1] required that the estimated sampling standard deviation s_{sam} should be less than 30 % of the target standard deviation σ_{p}, that is,

$$s_{\mathrm{sam}} < \sigma_{\mathrm{all}}$$

where the allowed sampling standard deviation $\sigma_{\mathrm{all}} = 0.3\sigma_{\mathrm{p}}$.

This condition, when fulfilled, was called "sufficient homogeneity". At that limit, the standard deviation of the resultant *z*-scores would be inflated by the heterogeneity by somewhat less than 5 % relative, for example, from 2.0 to 2.1, which was deemed to be acceptable. If the condition were not fulfilled, the *z*-scores would reflect, to an unacceptable degree, variation in the material as well as variation in laboratory performance. Participants in proficiency testing schemes need to be reassured that the distributed units of the test material are sufficiently similar, and this requirement usually calls for testing.

The test specified called for the selection of ten or more units at random after the putative homogenized material has been split and packaged into discrete samples for distribution. The material from each sample was then analyzed in duplicate, under randomized repeatability conditions (that is, all in one run) using a method with sufficient analytical precision. The value of σ_{sam} is then estimated from the mean squares after one-way analysis of variance (ANOVA).

Tests for sufficient homogeneity are in practice never wholly satisfactory. The main problem is that, because of the high cost of the analysis, the number of samples taken for testing will be small. This makes the power of the statistical test (that is, the probability of rejecting the material when it is in fact heterogeneous) relatively low. A further problem is that heterogeneity is inherently likely to be patchy, and discrepant distribution units might be under-represented among those selected for test. Homogeneity tests should be regarded as essential, but not foolproof, safeguards.

However, given that sufficient homogeneity is a reasonable prior assumption (because proficiency testing scheme providers do their best to ensure it), and that the cost of testing for it is often high, it is sensible to make the main emphasis the avoidance of "Type 1 errors" (that is, false rejection of a satisfactory material). That action would have the effect of making a modified test less prone to rejecting good samples.

To test for sufficient homogeneity, we have to estimate σ_{sam} from the results of a randomized replicated experiment by using ANOVA. In the experiment, each selected distribution unit is separately homogenized and analyzed in duplicate. Much depends on the quality of the analytical results. If the analytical method is sufficiently precise, σ_{sam} can be reliably estimated, and any lack of sufficient homogeneity present detected with reasonably high probability. In fact, the test could be too sensitive. The material can be significantly heterogeneous statistically, but the sampling variance negligible in relation to σ_{p}. However, if the analytical standard deviation σ_{an} is not small, important sampling variation may be obscured by analytical variation. We may get a nonsignificant result when testing for heterogeneity, not because it is not present, but because the test has no power to detect it.

The 1993 Harmonized Protocol, while specifying a need for sufficiently good analytical precision, did not specify any numerical limits on σ_{an}, but the above discussion suggests that it is desirable to do so. In setting this value, there has to be a trade-off between the cost of specifying very precise analytical methods and the risk of failing to detect important sampling variation. We accordingly recommend that the analytical (repeatability) precision of the method used in the homogeneity test should satisfy

$$\sigma_{\mathrm{an}}/\sigma_{\mathrm{p}} < 0.5$$

However, we have to recognize that occasionally it may impracticable to meet this requirement, hence the need for a statistical procedure that will give a sensible result regardless of the value of σ_{an}.

Recommendation 7

The analytical (repeatability) precision of the method used in the homogeneity test should satisfy $\sigma_{\mathrm{an}}/\sigma_{\mathrm{p}} < 0.5$ where σ_{an} is the repeatability standard deviation appropriate to the homogeneity test.

3.11.2 The new statistical procedure

Rather than express the criterion for sufficient homogeneity in terms of the estimated sampling variance s^2_{sam}, as in the 1993 Harmonized Protocol, it is more logical to impose a limit on the true sampling variance σ^2_{sam} [19]. It is this true sampling variance that is more relevant to the variability in the (untested) samples sent out to laboratories. Thus, our new criterion for sufficient homogeneity is that the sampling variance σ^2_{sam} must not exceed an allowable quantity $\sigma^2_{\mathrm{all}} = 0.09\sigma^2_{\mathrm{p}}$ (that is, $\sigma_{\mathrm{all}} = 0.3\sigma_{\mathrm{p}}$). Then in testing for homogeneity it makes sense to test the hypothesis $\sigma^2_{\mathrm{sam}} \le \sigma^2_{\mathrm{all}}$ against the alternative $\sigma^2_{\mathrm{sam}} > \sigma^2_{\mathrm{all}}$. (The usual *F*-test in a one-way ANOVA tests the rather stricter hypothesis $\sigma^2_{\mathrm{sam}} = 0$ against the alternative $\sigma^2_{\mathrm{sam}} > 0$, which would provide evidence that there is sampling variation, but not necessarily that it is unacceptably large.) The new procedure is designed to accommodate this requirement and the other difficulties referred to above. The complete procedure and a worked example are shown in Appendix 1.

Recommendation 8

Employ an explicit test of the hypothesis H: $\sigma^2_{\mathrm{sam}} \le \sigma^2_{\mathrm{all}}$, by finding a one-sided 95 % confidence interval for σ^2_{sam} and rejecting H when this interval does not include σ^2_{sam}. This is equivalent to rejecting H when

$$s^2_{\mathrm{sam}} > \mathrm{F}_1\sigma^2_{\mathrm{all}} + \mathrm{F}_2 s^2_{\mathrm{an}}$$

where s^2_{sam} and s^2_{an} are the usual estimates of sampling and analytical variances obtained from the ANOVA, and F_1 and F_2 are constants that may be derived from standard statistical tables.

3.11.3 Handling outlying results in homogeneity tests

Sporadic analytical outliers affect homogeneity-test data sets quite often, as at least 20 analytical results are produced in each test. Analytical outliers are manifested as an unexpectedly large deviation between the duplicated results on one of the samples. Regardless of the heterogeneity or otherwise of the original bulk material, as the procedure requires that each sample is properly homogenized before the two test portions are removed from it, any outlying difference between duplicate pairs must be due to the analysis rather than the material.

The effect of a single (that is, analytical) outlying result is perhaps unexpected: Although it inflates the estimate of the between-sample variance, an outlier helps the material pass the *F*-test because it inflates the estimate of analytical variance to a greater degree. The more extreme the analytical outlier, the closer the *F*-value becomes to unity (other results remaining equal). Thus, although the 1993 Harmonized Protocol called for all results to be retained, there is a clear case for excluding extreme analytical outliers when they can be unequivocally identified. If, however, a data set apparently contains more than one pair of discordant analytical results, the validity of the whole exercise is thrown into doubt and the homogeneity test data should be discarded.

Note that the recommendation to reject a single outlying pair only applies to samples with individual outlying results, not to samples with mutually consistent results but outlying means. If the results from a sample are concordant with one another, but the mean result is discordant with the other data, the results must be retained—they comprise evidence for between-sample heterogeneity. This distinction is illustrated in Fig. 1. Sample 9 provided discordant results that should be excluded. Sample 12 provided concordant results with an outlying mean, and must not be excluded. Cochran's variance test is suitable for detecting extreme differences between observations. (Appendix 1).

Note 1: Automatic rejection of variance outliers at the 95 % confidence level will materially increase the fraction of incorrect homogeneity test failures and is not recommended.

Note 2: In certain rare circumstances, the recommendation to discard a single pair of analytical outliers may be inappropriate. This exception can occur when the analyte is present in the test material at a low concentration overall, but is almost wholly confined to a trace phase that resists comminution but contains the analyte at high concentration. An example is a rock containing gold. Whether outlier rejection should be used is a matter for a scheme's advisory committee to determine, taking the above discussion into account.

Recommendation 9

In testing for sufficient homogeneity, duplicate results from a single distribution unit should be deleted before the analysis of variance if they are shown to be significantly different from each other by Cochran's test at the 99 % level of confidence or an equivalent test for extreme within-group variance. Data sets containing discrepancies in two such distribution units should be discarded *in toto*. Pairs of results with outlying mean value but no evidence of extreme variance should not be discarded.

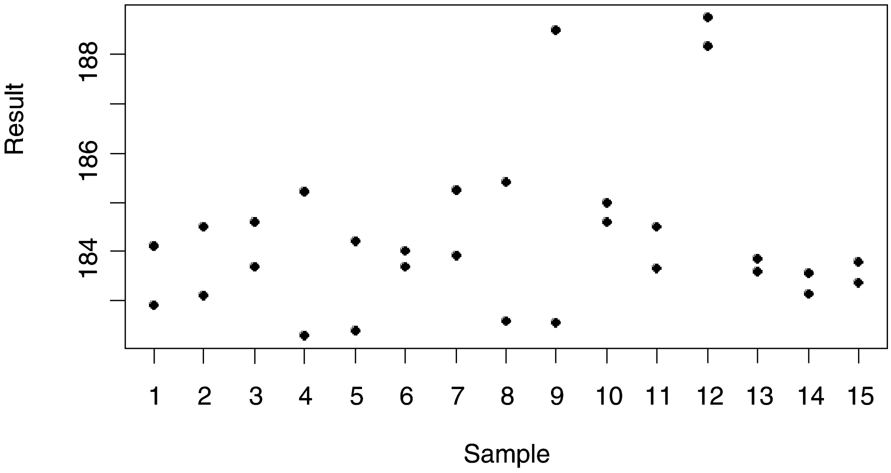


Fig. 1 Homogeneity test data.

3.11.4 Other pathologies of homogeneity test data sets

All aspects of testing for sufficient homogeneity depend on the laboratory carrying out the test correctly and, in particular, selecting the samples for test at random, homogenizing them before analysis, analyzing the duplicated test portions under strictly randomized conditions, and recording the results with sufficient digit resolution to allow the analysis of the variation. Any infringements may invalidate the outcome of the test. Unless strict control is maintained, data sets where at least some of these requirements have not been met are commonly encountered. We therefore recommend that (a) detailed instructions be issued to the laboratory conducting the homogeneity test, and (b) the data be checked for suspect features as a matter of routine. Suggestions for these instructions and tests are given in Appendix 1.

Recommendation 10

a. Detailed instructions should be issued to the laboratory conducting the homogeneity test.
b. The resultant data should be checked for suspect features.

3.11.5 Stability of test materials

Materials distributed in proficiency tests must be sufficiently stable over the period in which the assigned value is to be valid. The term "sufficiently stable" implies that any changes that occur during the relevant period must be of inconsequential magnitude in relation to the interpretation of the results of a round. Thus, if it were deemed that a change in the z-score of ±1 would be inconsequential, then an instability amounting to a change in analyte concentration of $0.1\sigma_{\mathrm{p}}$ could be tolerated. Normally, the period in question is the interval between preparation of the material and the deadline for return of the results, although the period will be longer if residual material is to be re-used in subsequent rounds or for other purposes. The stability test should involve exposure to the most extreme conditions likely to be encountered during the distribution and storage of the material, or to accelerated degradation conditions. The material under test should be in the packaging in which it is to be distributed.

A comprehensive test for sufficient stability would be extremely demanding of resources (see below). It is therefore not usually practicable to test every batch of material for every round in a series. However, it is a sensible prior precaution to test each new combination of material and analyte before it is first used in a proficiency test and occasionally thereafter, and the following paragraphs discuss this. It may additionally be useful to monitor stability by, for example, arranging for analysis of units pre- and post-distribution by a single laboratory, providing for return of some distributed units for direct comparison with stored units, or comparing post-distribution analysis results with prior information such as homogeneity test data.

Basic stability tests involve a comparison of the apparent analyte levels between material subjected to likely decomposition conditions and material which has not been so treated. This usually requires a sample of the distribution units to be randomly divided into (at least) two equal subsets. The "experimental" subset is subjected to the appropriate treatment, while the "control" subset is kept under conditions of maximum stability, for example, at low temperatures and low oxygen tension. Alternatively, and especially if stability for extended periods is of interest, the control subset may be kept under ambient conditions while the experimental subset is kept under conditions of accelerated decomposition (e.g., higher temperatures). The materials are then analyzed simultaneously, or if that is impossible, as a randomized block design.

Such experiments must be carefully designed to avoid compounding the effects of change in the material with variation in the efficacy of the analytical method used. Analysis of the control material at the beginning of the test period and experimental material at the end automatically includes any run-to-run analytical difference in the results and may well lead to the incorrect conclusion that there is a significant instability. The recommended approach is, if at all possible, to analyze the experimental and control subsets together, in a random order within a single run of analysis, that is, under repeatability conditions. Any highly significant difference between the mean results of the two subsets can then safely be regarded as evidence of instability.

As in homogeneity testing, a conceptual distinction must be made between statistically significant instability and consequential instability. For instance, a highly significant change in the analytical results might be detected, but the change may still be so small that a negligible effect on the z-scores of the participants could be inferred. In practice, significance tests are not powerful enough to validate such a small instability unless an exceptionally precise analytical method is used, and/or inordinate numbers of distribution units analyzed. The stability test will, therefore, only detect a gross instability.

COLLECTED RECOMMENDATIONS

Recommendation 1: Scheme for obtaining a consensus value and its uncertainty (see Section 3.3.2)

a. Exclude from the data any results that are identifiably invalid (e.g., if they are expressed in the wrong units or obtained by using a proscribed method) or are extreme outliers (e.g., outside the range of ±50 % of the median).

b. Examine a visual presentation of the remaining results, by means of a dot plot [for small ($n < 50$) data sets], bar chart, or histogram (for larger data sets). If outliers cause the presentation of the bulk of the results to be unduly compressed, make a new plot with the outliers deleted. If the distribution is, outliers aside, apparently unimodal and roughly symmetric, go to (c), otherwise go to (d).

c. Calculate the robust mean $\hat{\mu}_{\mathrm{rob}}$ and standard deviation $\hat{\sigma}_{\mathrm{rob}}$ of the n results. If $\hat{\sigma}_{\mathrm{rob}}$ is less than about $1.2\sigma_{\mathrm{p}}$, then use $\hat{\mu}_{\mathrm{rob}}$ as the assigned value x_{a} and $\hat{\sigma}_{\mathrm{rob}}/\sqrt{n}$ as its standard uncertainty. If $\hat{\sigma}_{\mathrm{rob}} > 1.2\sigma_{\mathrm{p}}$, go to (d).

d. Make a kernel density estimate of the distribution of the results using normal kernels with a bandwidth h of $0.75\sigma_{\mathrm{p}}$. If this results in a unimodal and roughly symmetric kernel density, and the mode and median are nearly coincident, then use $\hat{\mu}_{\mathrm{rob}}$ as the assigned value x_{a} and $\hat{\sigma}_{\mathrm{rob}}/\sqrt{n}$ as its standard uncertainty. Otherwise, go to (e).

e. If the minor modes can safely be attributed to outlying results, and are contributing less than about 5 % to the total area, then still use $\hat{\mu}_{\mathrm{rob}}$ as the assigned value, x_{a} and $\hat{\sigma}_{\mathrm{rob}}/\sqrt{n}$ as its standard uncertainty. Otherwise, go to (f).

f. If the minor modes make a considerable contribution to the area of the kernel, consider the possibility that two or more discrepant populations are represented in the participants' results. If it is possible to infer from independent information (e.g., details of the participants' analytical methods) that one of these modes is correct and the others are incorrect, use the selected mode as the assigned value x_{a} and its standard error as its standard uncertainty. Otherwise, go to (g).

g. The methods above having failed, abandon the attempt to determine a consensus value and report no individual laboratory performance scores for the round. It may still be useful, however, to provide the participants with summary statistics on the data set as a whole.

Recommendation 2: Using uncertainty on the assigned value (see Section 3.4)
The proficiency test provider should nominate a multiplier $0.1 < l < 0.5$ appropriate for the scheme and, having evaluated $u^2(x_{\mathrm{a}})$ for a round, act as follows:

- if $u^2(x_{\mathrm{a}})/\sigma_{\mathrm{p}}^2 \le 0.1$, issue unqualified z-scores;
- if $0.1 < u^2(x_{\mathrm{a}})/\sigma_{\mathrm{p}}^2 \le l$, issue qualified z-scores (such as "provisional z-scores"); and
- if $u^2(x_{\mathrm{a}})/\sigma_{\mathrm{p}}^2\ l$ do not issue z-scores.

Note: In the inequality $0.1 < l < 0.5$, the limits can be modified somewhat to meet the exact requirements of particular schemes.

Recommendation 3: Determination of the standard deviation for proficiency assessment (see Section 3.5)
Wherever possible, the proficiency testing scheme should use for σ_{p}, the standard deviation for proficiency assessment, a value that reflects fitness-for-purpose for the sector. If there is no single level that is generally appropriate, the provider should refrain from calculating scores, or should show clearly on the reports that the scores are for informal descriptive use only and not to be regarded as an index of performance of the participants.

Recommendation 4: Use of weighting in calculating consensus values (see Section 3.6)
Even when uncertainty estimates are available, unweighted methods (i.e., taking no account of the individual uncertainties) should be used to obtain the consensus value and its uncertainty, according to the methods described in Sections 3.3 and 3.4.

Recommendation 5: Scoring results that are reported with uncertainty (see Section 3.6.2)
Schemes should not provide zeta scores unless there are special reasons for doing so. Where a participant has requirements inconsistent with that of the scheme, the participant may calculate zeta scores or the equivalent.

Recommendation 6: Classification and ranking of laboratories (see Section 3.9)
Proficiency scheme providers, participants and end-users should avoid classification and ranking of laboratories on the basis of their *z*-scores.

Recommendation 7: Repeatability requirement in homogeneity testing (see Section 3.11.1)
The analytical (repeatability) precision of the method used in the homogeneity test should satisfy $\sigma_{\mathrm{an}}/\sigma_{\mathrm{p}} < 0.5$ where σ_{an} is the repeatability standard deviation appropriate to the homogeneity test.

Recommendation 8: Statistical test in homogeneity testing (see Section 3.11.2)
Employ an explicit test of the hypothesis H: $\sigma^2_{\mathrm{sam}} \le \sigma^2_{\mathrm{all}}$, by finding a one-sided 95 % confidence interval for σ^2_{sam} and rejecting H when this interval does not include σ^2_{all}. This is equivalent to rejecting H when

$$s^2_{\mathrm{sam}} > \mathrm{F}_1\sigma^2_{\mathrm{all}} + \mathrm{F}_2 s^2_{\mathrm{an}} \qquad (6)$$

where s^2_{sam} and s^2_{an} are the usual estimates of sampling and analytical variances obtained from the ANOVA, and F_1 and F_2 are constants that may be derived from standard statistical tables.

Recommendation 9: Handling outliers in homogeneity testing (see Section 3.11.3)
In testing for sufficient homogeneity, duplicate results from a single distribution unit should be deleted before the analysis of variance if they are shown to be significantly different from each other by Cochran's test at the 99 % level of confidence or an equivalent test for extreme within-group variance. Data sets containing discrepancies in two such distribution units should be discarded *in toto*. Pairs of results with outlying mean value but no evidence of extreme variance should not be discarded.

Recommendation 10: Management of homogeneity tests (see Section 3.11.4)
a. Detailed instructions should be issued to the laboratory conducting the homogeneity test.
b. The resultant data should be checked for suspect features.

REFERENCES

1. M. Thompson and R. Wood. "The International Harmonised Protocol for the proficiency testing of (chemical) analytical laboratories", *Pure Appl. Chem.* **65**, 2123–2144 (1993). [Also published in *J. AOAC Int.* **76**, 926–940 (1993)].
2. See: (a) M. Golze. "Information system and qualifying criteria for proficiency testing schemes", *Accred. Qual. Assur.* **6**, 199–202 (2001); (b) J. M. F. Nogueira, C. A. Nieto-de-Castro, L. Cortez. "EPTIS: The new European database of proficiency testing schemes for analytical laboratories", *J. Trends Anal. Chem.* **20**, 457–61 (2001); (c) <http://www.eptis.bam.de>.
3. R. E. Lawn, M. Thompson, R. F. Walker. *Proficiency Testing in Analytical Chemistry*, The Royal Society of Chemistry, Cambridge (1997).
4. International Organization for Standardization. *ISO Guide 43: Proficiency testing by interlaboratory comparisons—Part 1: Development and operation of proficiency testing schemes*, Geneva, Switzerland (1994).

5. International Organization for Standardization. *ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparisons*, Geneva, Switzerland (2005).
6. ILAC-G13:2000. *Guidelines for the requirements for the competence of providers of proficiency testing schemes*. Available online at <http://www.ilac.org/>.
7. M. Thompson, S. L. R. Ellison, R. Wood. "Harmonized guidelines for single laboratory validation of methods of analysis", *Pure Appl. Chem.* **74**, 835–855 (2002).
8. International Organization for Standardization. *ISO Guide 33:2000, Uses of Certified Reference Materials*, Geneva, Switzerland (2000).
9. M. Thompson and R. Wood. "Harmonised guidelines for internal quality control in analytical chemistry laboratories", *Pure Appl. Chem.* **67**, 649–666 (1995).
10. T. Fearn, S. Fisher, M. Thompson, S. L. R. Ellison. "A decision theory approach to fitness-for-purpose in analytical measurement", *Analyst* **127**, 818–824 (2002).
11. Analytical Methods Committee. "Robust statistics—how not to reject outliers: Part 1 Basic concepts", *Analyst* **114**, 1693 (1989).
12. M. Thompson. "Bump-hunting for the proficiency tester: Searching for multimodality", *Analyst* **127**,1359–1364 (2002).
13. M. Thompson. "A natural history of analytical methods", *Analyst* **124**, 991 (1999).
14. M. Thompson. "Recent trends in interlaboratory precision at ppb and sub-ppb concentrations in relation to fitness-for-purpose criteria in proficiency testing", *Analyst* **125**, 385–386 (2000).
15. Analytical Methods Committee. "Uncertainty of measurement—implications for its use in analytical science", *Analyst* **120**, 2303–2308 (1995).
16. W. P. Cofino, D. E. Wells, F. Ariese, J.-W. M. Wegener, R. I. H. M. Stokkum, A. L. Peerboom. "A new model for the inference of population characteristics from experimental data using uncertainties. Application to interlaboratory studies", *Chemom. Intell. Lab Systems* **53**, 37–55 (2000).
17. T. Fearn. "Comments on 'Cofino Statistics'", *Accred. Qual. Assur.* **9**, 441–444 (2004).
18. M. Thompson and P. J. Lowthian. "The frequency of rounds in a proficiency test: does it affect the performance of participants?", *Analyst* **123**, 2809–2812 (1998).
19. T. Fearn and M. Thompson. "A new test for 'sufficient homogeneity'", *Analyst* **126**, 1414–1417 (2001).

APPENDIX 1: RECOMMENDED PROCEDURE FOR TESTING A MATERIAL FOR SUFFICIENT HOMOGENEITY

A1.1 Procedure

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 of 10 containers strictly at random.
4. Separately homogenize the contents of each of the m selected containers and take two test portions from each.
5. Analyze the $2m$ 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_{sam}. If possible, $\sigma_{\mathrm{an}} < 0.5\sigma_{\mathrm{p}}$.

The first step is to examine the data for pathologies. Such a check could be made visually on a simple plot of the results vs. sample number, searching for such diagnostic features as: (a) trends or discontinuities; (b) nonrandom distribution of differences between first and second test results; (c) excessive rounding; and (d) outlying results within samples.

If all is well, we use the data to estimate the analytical and sampling variances. If a program to perform a one-way analysis of variance is available, this may be used. Alternatively, a full calculation scheme is given below.

A1.2 Statistical methods

A1.2.1 Cochran test procedure for duplicate results

Calculate the sum, S_i, and difference, D_i, of each pair of duplicates, for $i = 1, ..., m$.

Calculate the sum of squares S_{DD} of the m differences from

$$S_{\mathrm{DD}} = \sum_{m} D_i^2$$

Cochran's test statistic is the ratio of $D^2_{\max}$, the largest squared difference to this sum of squared differences

$$C = \frac{D^2_{\max}}{S_{\mathrm{DD}}}$$

Calculate the ratio and compare it with the appropriate critical value from tables. Table 1 gives values of the critical values for 95 and 99 % confidence for m between 7 and 20 pairs.

Results for Cochran outlying pairs detected at the 95 % or higher level of confidence should always be inspected closely for evidence of transcription or other errors in the analysis and appropriate action taken if any errors are found. An outlying pair should not be rejected unless it is significant at the 99 % level or irremediable analytical procedure errors are found. A single Cochran outlier at the 99 % level should be excluded from the ANOVA unless there is reason to the contrary (see Section 3.11).

A1.2.2 Test for significant inhomogeneity

Use the same sum of squared differences to calculate

$$s_{\mathrm{an}}^2 = \Sigma D_i^2/2m$$

Calculate the variance V_S of the sums S_i

$$V_\mathrm{S} = \Sigma(S_i - \bar{S})^2/(m - 1)$$

where $\bar{S} = (1/m)\Sigma S_\mathrm{i}$ is the mean of the S_i.

Calculate the sampling variance ${s_\mathrm{sam}}^2$ as

$${s_\mathrm{sam}}^2 = (V_\mathrm{S}/2 - s_\mathrm{an})/2$$

or as ${s_\mathrm{sam}}^2 = 0$ if the above estimate is negative.

If a program for one-way analysis of variance is available, the quantities $V_\mathrm{S}/2$ and s_an above may be extracted from the analysis of variance table as the "between" and "within" mean squares, respectively.

Calculate the allowable sampling variance σ_{all}^2 as

$$\sigma_{\mathrm{all}}^2 = (0.3\sigma_\mathrm{p})^2$$

where σ_p is the standard deviation for proficiency assessment.

Taking the values of F_1 and F_2 from Table 2, calculate the critical value for the test as $c = F_1\sigma_{\mathrm{all}}^2 + F_2 s_{\mathrm{an}}^2$

If ${s_\mathrm{sam}}^2 > c$, there is evidence (significant at the 95 % level of confidence) that the sampling standard deviation in the population of samples exceeds the allowable fraction of the target standard deviation, and the test for homogeneity has failed.

If ${s_\mathrm{sam}}^2 < c$, there is no such evidence, and the test for homogeneity has been passed.

A1.2.3 Tables of critical values for homogeneity testing

Table 1 Critical values for the Cochran test statistic for duplicates.

m	95 %	99 %
7	0.727	0.838
8	0.68	0.794
9	0.638	0.754
10	0.602	0.718
11	0.57	0.684
12	0.541	0.653
13	0.515	0.624
14	0.492	0.599
15	0.471	0.575
16	0.452	0.553
17	0.434	0.532
18	0.418	0.514
19	0.403	0.496
20	0.389	0.480

Table 2 Factors F_1 and F_2 for use in testing for sufficient homogeneity.

m*	20	19	18	17	16	15	14	13	12	11	10	9	8	7
F_1	1.59	1.60	1.62	1.64	1.67	1.69	1.72	1.75	1.79	1.83	1.88	1.94	2.01	2.10
F_2	0.57	0.59	0.62	0.64	0.68	0.71	0.75	0.80	0.86	0.93	1.01	1.11	1.25	1.43

*m is the number of samples that have been measured in duplicate.

The two constants in Table 2 are derived from standard statistical tables as $F_1 = \chi^2_{m-1,0.95}/(m-1)$, where $\chi^2_{m-1,0.95}$ is the value exceeded with probability 0.05 by a chi-squared random variable with $m - 1$ degrees of freedom, and $F_2 = (F_{m-1,m,0.95} - 1)/2$ where $F_{m-1,m,0.95}$ is the value exceeded with probability 0.05 by a random variable with an F-distribution with $m - 1$ and m degrees of freedom.

A1.3 Example instructions for laboratories testing for sufficient homogeneity

The laboratory should be experienced in the analytical method used.

- Select the $m \geq 10$ distribution units strictly at random from the complete set. This must be done in a formal way, by assigning a sequential number to the units, either explicitly (by labeling them), or implicitly (e.g., by their position in an array). Make the selection by use of random numbers from a table or generated (with a new seed each time) by a computer package (e.g., Microsoft Excel). It is not acceptable to select the units in any other way (e.g., by shuffling them). Do not use a random sequence previously used.
- Homogenize the contents of each distribution unit in an appropriate manner (e.g., in a blender) and from each weigh out two test portions. Label the test portions as shown.

Sequential code of distribution unit	Label of first test portion	Label of second test portion
1	1.1	1.2
2	2.1	2.2
3	3.1	3.2
.	.	.
.	.	.
m	m.1	m.2

- Sort the 20 test portions into a random order and carry out all analytical operations on them in that order. Again, random number tables or a computer package must be used to generate a new random sequence. An example random sequence (not to be copied) is 7.1 3.1 5.2 5.1 10.2 1.1 2.1 9.2 8.2 1.2 4.1 2.2 9.1 10.1 7.2 3.2 8.1 6.1 4.2 6.2
- Conduct the analysis if at all possible under repeatability conditions (i.e., in one run) or, if that is impossible, in successive runs with as little change as possible to the analytical system, using a method that has a repeatability standard deviation of less than $0.5\sigma_p$. Record results if possible with as many significant figures as those required for $0.01\sigma_p$.
- Return the 20 analytical results, including the labels, in the run order used.

A1.4 Homogeneity testing: Example

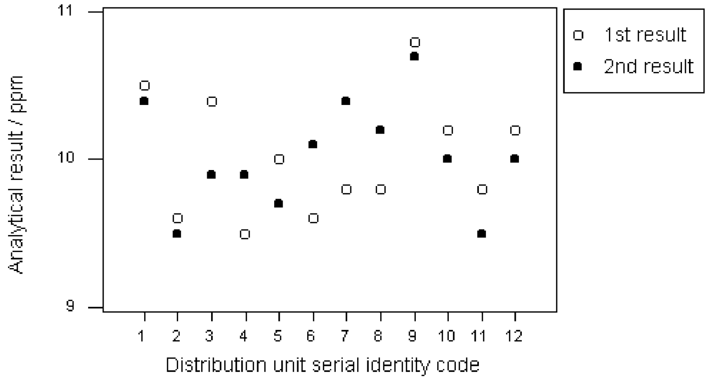
A1.4.1 The data

The data, shown in Table 3, are taken from the 1993 Harmonized Protocol [Ref. 1 of Part 3].

Table 3 Duplicated results for 12 distribution units of soya flour analyzed for copper (ppm), together with some intermediate stages of the ANOVA calculation.

Sample	Result *a*	Result *b*	$D = a - b$	$S = a + b$	$D^2 = (a - b)^2$
1	10.5	10.4	0.1	20.9	0.01
2	9.6	9.5	0.1	19.1	0.01
3	10.4	9.9	0.5	20.3	0.25
4	9.5	9.9	–0.4	19.4	0.16
5	10.0	9.7	0.3	19.7	0.09
6	9.6	10.1	–0.5	19.7	0.25
7	9.8	10.4	–0.6	20.2	0.36
8	9.8	10.2	–0.4	20.0	0.16
9	10.8	10.7	0.1	21.5	0.01
10	10.2	10.0	0.2	20.2	0.04
11	9.8	9.5	0.3	19.3	0.09
12	10.2	10.0	0.2	20.2	0.04

A1.4.2 Visual appraisal



The data are presented visually above, and show no suspect features such as discordant duplicated results, outlying samples, trends, discontinuities, or any other systematic effects. (A Youden plot, of first vs. second duplicate, could also be used.)

A1.4.3 Cochran's test

The largest value of D^2 is 0.36 and the sum of D^2 is 1.47, so the Cochran test statistic is 0.36/1.47 = 0.24. This is less than the 5 % critical value of 0.54, so there is no evidence for analytical outliers and we proceed with the complete data set.

A1.4.4 Homogeneity test

Analytical variance: s^2_{an} =1.47/24 = 0.061

Between-sample variance: The variance of the sums $S = a + b$ is 0.463, so

s^2_{sam}= (0.463/2 – 0.061)/2 = (0.231 – 0.061)/2 = 0.085

Acceptable between-sample variance: The target standard deviation is 1.14 ppm, so the allowable between-sample variance is σ^2_{all} = (0.3 × 1.14)2 = 0.116.

Critical value: The critical value for the test is 1.79 σ^2_{all} + 0.86 s^2_{an} = 1.79 × 0.116 + 0.86 × 0.061 = 0.26.

Since s^2_{sam} = 0.085 < 0.26, passed and the material is sufficiently homogeneous.

APPENDIX 2: EXAMPLE OF CONDUCTING A TEST FOR STABILITY

The procedure outlined in Section 3.11.5 has been carried out. The standard deviation for proficiency assessment (σ_p) has been set at 0.1*c* (i.e., an RSD of 10 %), and the results of the analysis under repeatability conditions, in the random order in which the analyses have been carried out, are as tabulated below.

Material	Result/ppm
Experimental	11.5
Control	13.4
Control	12.2
Experimental	12.3
Control	12.7
Experimental	10.9
Control	12.5
Experimental	11.4
Experimental	12.4
Control	12.5

A two-sample *t*-test with pooled standard deviation gives the following statistics:

	n	$\bar{x}$
Control	5	12.66
Experimental	5	11.70
Difference		0.96

Pooled standard deviation: 0.551

95 % confidence interval for ($\mu_{\text{cont}} - \mu_{\text{expt}}$): (0.16, 1.76)

The *t*-test of H_0: $\mu_{\text{cont}} = \mu_{\text{expt}}$ vs H_A: $\mu_{\text{cont}} \neq \mu_{\text{expt}}$ gives a value of $t = 2.75$ with 8 degrees of freedom, corresponding with a probability (*p*-value) of 0.025. The instability difference of 0.96 ppm is, therefore, significant at the 95 % level of confidence. (This can also be deduced from the 95 % confidence interval, which does not include zero.)

Using the mean of about 12 as the concentration of the analyte, we find σ_p = 0.1 × 12 = 1.2. The difference due to instability is much greater than the desired limit of 0.1σ_p, so there is consequential instability and the material is unsuitable for use.

APPENDIX 3: EXAMPLES OF PRACTICE IN DETERMINING A PARTICIPANT CONSENSUS FOR USE AS AN ASSIGNED VALUE

A3.1 Example 1

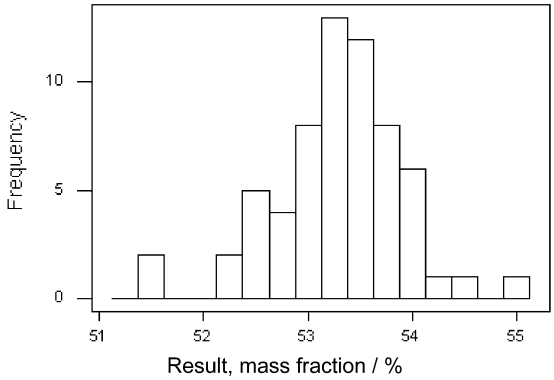
Participants' results are listed in the following table, which also shows the relevant summary statistics. The units are mass fraction, expressed as a percentage (%); numerical precision is as reported by participants.

Reported results						
54.09	53.15	53.702	52.9	53.65	52.815	53.5
52.95	52.35	53.49	55.02	53.32	54.04	53.15
53.41	53.4	53.3	54.33	52.83	53.4	53.38
53.19	52.4	52.9	53.44	53.75	53.39	53.661
54.09	53.09	53.21	53.12	53.18	53.3	52.62
53.7	53.51	53.294	53.57	52.44	53.04	53.23
63.54	46.1	53.18	54.54	53.76	54.04	53.64
53	54.1	52.2	52.54	53.42	53.952	50.09
53.06	48.07	52.51	51.44	52.72	53.7	
53.16	53.54	53.37	51.52	46.85	52.68	

Summary statistics	
n	68
Mean	53.10
Standard deviation.	1.96
Median	53.30
H15 estimate of the mean*	53.24
H15 estimate of standard deviation*	0.64

*See refs. [1] and [2] for this Appendix.

The standard deviation for proficiency assessment σ_{p} is 0.6 %



The histogram (outliers aside) suggests a unimodal and roughly symmetric distribution.

The summary statistics show an almost coincident robust mean and median. The robust standard deviation is less than $1.2\sigma_p$, so there is no concern about wide distributions. The value of $\hat{\sigma}_{rob}/\sqrt{n}$ = 0.079, which is well below the guideline of $0.3\sigma_p = 0.17$.

The consensus value and its standard uncertainty are taken as mass fractions 53.24 and 0.08 %, respectively.

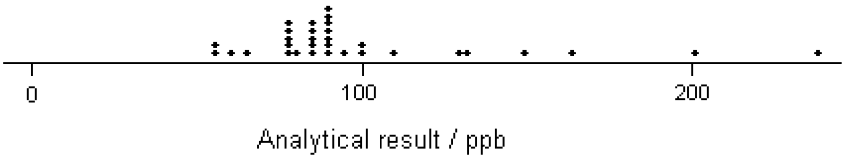
A3.2 Example 2

Participants reported the following results (units are ppb, that is, 10^9 mass fraction):

Reported results/ppb						
133	89	55	84.48	84.4	90.4	66.6
77	80	60.3	84	78	85	130
90	79	99.7	149	91	164	
78	84	110	77	91	89	
95	55	90	100	200.56	237	

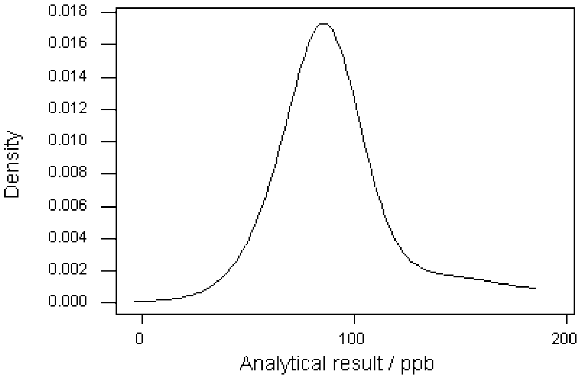
Summary statistics	
n	32
Mean	99.26
Standard deviation	39.76
Median	89.0
H15 estimate of the mean	91.45
H15 estimate of standard deviation	23.64

A dotplot (below) of the reported results shows a data set with a strong positive skew, which might cast doubt on the validity of robust statistics.



A provisional standard deviation for proficiency assessment, derived from the robust mean, was given by the Horwitz function: $\sigma_p = 0.452 \times 91.4^{\,0.8495} = 20.8$ ppb.

In view of the skew and the high robust standard deviation, the robust mean was suspect, so a kernel density distribution was constructed with a bandwidth h of $0.75\sigma_p$:



The kernel density shows a single mode at 85.2 ppb, and bootstrapping the data provided a standard error for the mode of 2.0 ppb. The revised σ_p based on a concentration of 85.2 is 19.7 ppb. The implied uncertainty of the mode (2.0) is below the guideline of $0.3\sigma_p = 5.9$ ppb.

The consensus and its standard uncertainty are taken as 85 and 2 ppb, respectively.

A3.3 Example 3

This is the first round of a series after the method of reporting had been changed, so as to quantify a different "weighing form". The ratio of relative molecular masses of the new and old weighing forms was 1.35.

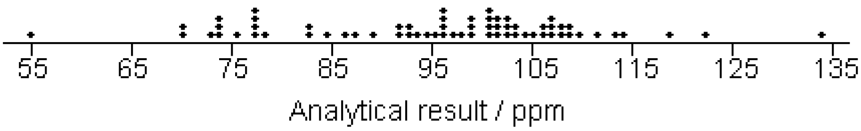
The standard deviation for proficiency assessment for the series is determined by the Horwitz function, namely $\sigma_p = 0.16 \times c^{\ 0.8495}$ where σ_p and c are in ppm (10^6 mass fraction).

The participants' results (in ppm, that is, 10^6 mass fraction) were:

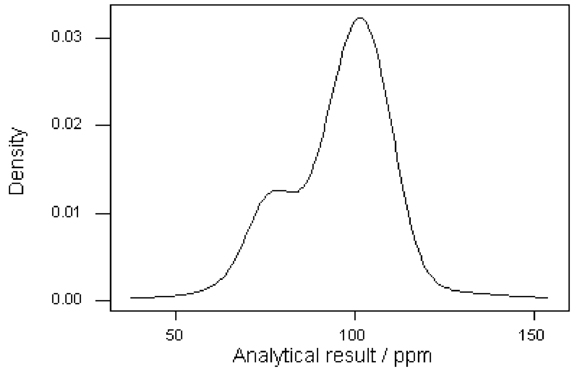
Reported results / ppm						
102.5	97.9	102	101	99	75.9	101
74	94	93	70	82.9	106	113
122	97	114	101	70	103.88	93
96	107	103	96	119	99	83
107	101	134	109	103.8	106	77
95	108	96	104	101.33	92.2	
94.5	102	77	98.91	107	109	
89	110	103	112	55	87	
108	105.4	86	74	73	77	
96	77.37	73.5	78	92	84.6	

Summary statistics	
n	65
Mean	95.69
Standard deviation	14.52
Median	98.91
H15 estimate of the mean	95.78
H15 estimate of standard deviation	14.63

As the dotplot (below) shows, this possibly represents a bimodal population.



The provisional value of σ_p derived from the robust mean is 7.71 ppm, but the robust standard deviation is considerably greater than that, so there are strong grounds for suspecting a mixed distribution. A kernel density was produced using a bandwidth h of 0.75 × 7.71 = 5.78.



This density function has modes at 78.6 and 101.5 ppm, with standard errors estimated by the bootstrap as 13.6 and 1.6, respectively. The ratio of the modes is 101.5/78.6 = 1.29, which is close to the ratio 1.35 of the relative molecular masses, which justifies the assumption that the major mode is correct and the minor mode represents participants who have incorrectly reported the old-style weighing form.

The consensus is, therefore, identified as 101.5 with an uncertainty of 1.6 ppm. The revised target value based on this consensus is 8.1. As the uncertainty of 1.6 is less than 0.3 × 8.1 = 2.43, the uncertainty is acceptable and the consensus can be used as the assigned value.

REFERENCES

1. Analytical Methods Committee. "Robust statistics—how not to reject outliers: Part 1 Basic concepts", *Analyst* **114**, 1693 (1989).
2. International Organization for Standardization. *ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparisons*, Geneva, Switzerland (2005).

APPENDIX 4: ASSESSING *Z*-SCORES IN THE LONGER TERM: SUMMARY SCORES AND GRAPHICAL METHODS

While a single z-score comprises a valuable indication of the performance of a laboratory, the consideration of a set or sequence of z-scores provides a deeper insight. Moreover, z-scores collected over time (for a single analyte/test material combination) reflect on the participant's uncertainty. Both graphical and statistical methods may be appropriate to examine collections of z-scores. However, in interpreting summary statistics, due caution is required to avoid incorrect conclusions. It is especially emphasized that use of a summary score derived from z-scores relating to a number of different analytes is not recommended; it has a very limited range of valid applications and tends to conceal sporadic or persistent problems with individual analytes. Moreover, it is prone to misuse by non-scientists.

A4.1 Summary scores

The following two types of summary score are statistically soundly based and may be useful for individual participants to assess a sequence of z-scores $[z_1, z_{2,} \dots z_i, \dots z_n]$ derived from a single combination of analyte, test material, and method.

The rescaled sum of the z-scores

$$S_{Z,\mathrm{rs}} = \sum_i z_i / \sqrt{n}$$

can be interpreted on the same basis as a single z-score, i.e., it is expected to be zero-centered with unit variance if the z-scores are. This statistic has the useful property of demonstrating a persistent bias or trend, so that a sequence of results [1.5, 1.5, 1.5, 1.5] would provide a statistically significant $S_{Z,\mathrm{rs}}$ of 3.0, even though any single one result is not significant. However, it could conceal two large z-scores of opposite sign, such as occur in the sequence [1.5, 4.5, –3.6, 0.6].

The sum of the squared z-scores

$$S_{ZZ} = \sum_i z_i^2$$

could be interpreted as a χ_n^2 distribution for zero-centered z-scores with unit variance. This statistic has the advantage of avoiding the cancellation of large z-scores of opposite sign, but is less sensitive to small biases.

Both of these summary statistics need to be protected (e.g., by robustification or filtering) against past outlying scores, which would otherwise have a long-term persistence. S_{ZZ} is especially sensitive to outliers. Both statistics (when so robustified) can be related to uncertainty of measurement in the following way. If the z-scores are based on fitness-for-purpose and therefore taken to be random $N(0,1)$, significantly high levels of the summary statistics indicate that the participant's uncertainty of measurement is greater than indicated by the schemes fitness-for-purpose criterion.

A4.2 Graphical methods

Graphical methods of summarizing a set of z-scores can be just as informative as summary scores and can be less prone to misinterpretation. Shewhart charts (with warning and action limits at $z = \pm 2$ and $z = \pm 3$, respectively) can be applied. Multiple univariate symbolic charts [1], such as those shown below, give a clear overview and are especially useful when scores from a group of analytes determined by a common method are considered. Hand-drawn charts are quick to update and serve just as well as those produced by computer.

The control chart (Fig. A4.1) shows upward-pointing symbols to indicate z-scores greater than zero and downward-pointing symbols for those less than zero. Small symbols represent instances where $2 \le |z| < 3$, and large symbols instances where $|z| \ge 3$ The data illustrated immediately show some noteworthy features. Results from round 11 are mostly too low, demonstrating a procedure that was faulty in some general feature, while analyte 7 gives high results too frequently, demonstrating a persistent problem with that specific analyte. The remaining results are roughly consistent with fitness-for-purpose, which on average would result in about 5 % of z-scores represented by small symbols.

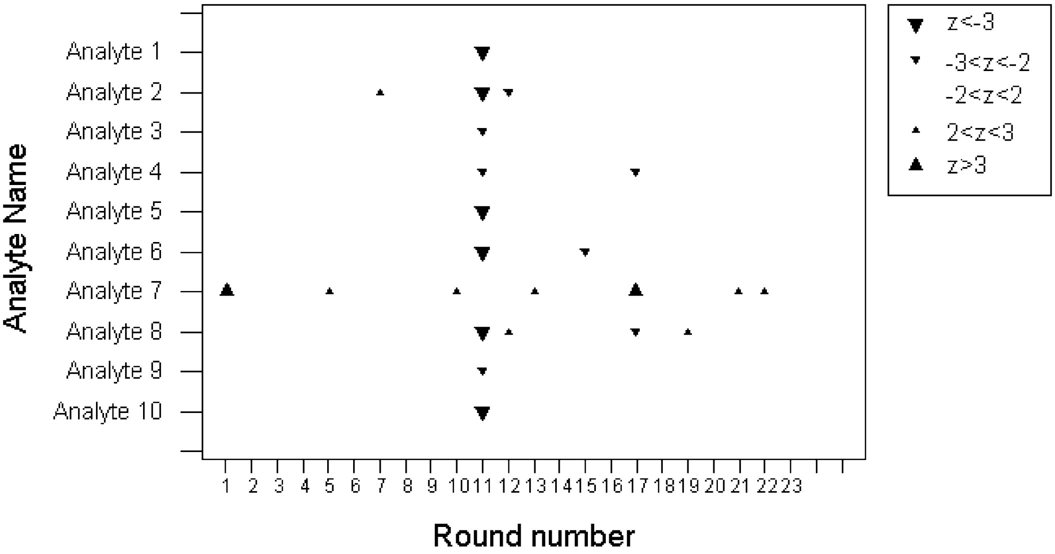


Fig A4.1 Control chart for *z*-scores.

A *J*-chart (otherwise known as a "zone chart") is even more informative, because it combines the capabilities of the Shewhart and the cusum charts. It does this by cumulating special *J*-scores attributed to successive results on either side of the zero line. This enables persistent minor biases to be detected as well as abrupt large changes in the analytical system. The usual rules for converting *z*-scores to J are as follows.

If	$z \geq 3$	then $J = 8$.
If	$2 \leq z < 3$	then $J = 4$.
If	$1 \leq z < 2$	then $J = 2$.
If	$-1 < z < 1$	then $J = 0$.
If	$-2 < z \leq -1$	then $J = -2$.
If	$-3 < z \leq -2$	then $J = -4$.
If	$z \leq -3$	then $J = -8$.

J-scores from successive rounds are cumulated until $|z| \geq 8$, which is an excursion beyond the action limits, and investigative procedures triggered. The cumulator is reset to zero immediately (i.e., before any further cumulation), after any such excursion and when successive values of *J* are of opposite sign.

Several examples of the cumulative effect of bias are visible in Fig. A4.2 (which illustrates the same results as Fig. A4.1 for comparison). For example, analyte 3 in rounds 1–4 receives *z*-scores of 1.5, 1.2, 1.5, and 1.1 respectively, translating into *J*-values of 2, 2, 2, and 2, which cumulate to 8 by round 4 and trigger investigative procedures. Similar examples are to be seen for analyte 7.

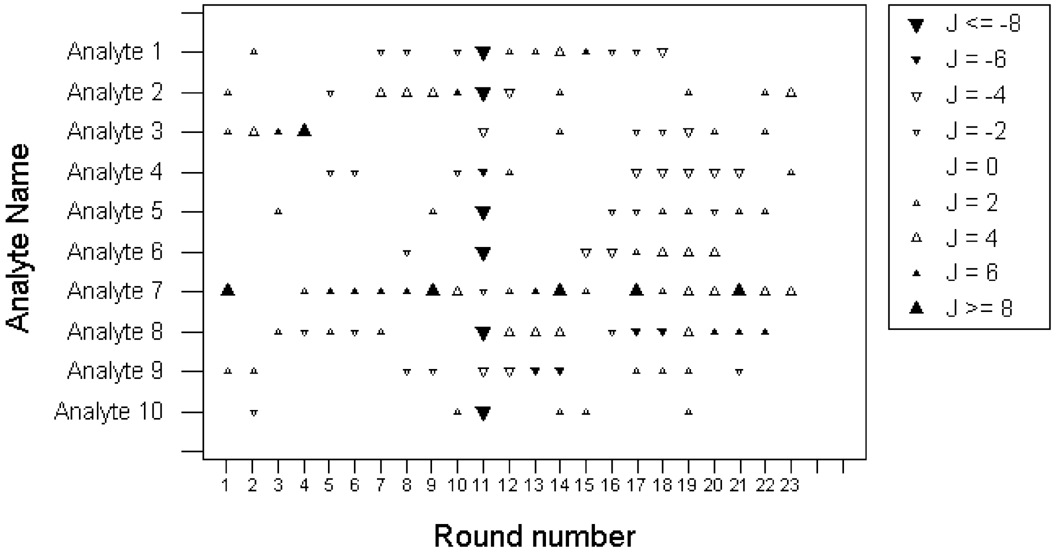


Fig. A4.2 J-chart for *z*-scores (same data as previous chart).

REFERENCES

1. M. Thompson, K. M. Malik, R. J. Howarth. "Multiple univariate symbolic control chart for internal quality control of analytical data", *Anal. Comm.* **35**, 205–208 (1998).
2. Analytical Methods Committee. "The J-chart: A simple plot that combines the capabilities of Shewhart and cusum charts, for use in analytical quality control", AMC Technical Briefs: No 12. <www.rsc.org/amc/>.

APPENDIX 5: METHOD VALIDATION THROUGH THE RESULTS OF PROFICIENCY TESTING SCHEMES

The purpose of a proficiency testing scheme is to test the accuracy of the participant laboratories. Participants have a free choice of method of analysis and commonly use a multiplicity of methods (or variants of a single "method"). Consequently, there is usually no scope for method validation as a by-product of proficiency testing. However, method validation becomes a possibility if there are sufficient participants in the proficiency testing scheme who use the same closely defined method of analysis. This possibility, if exploited properly, can be seen as an inexpensive alternative to, or a confirmation of, the collaborative trial, which is the recognized [1] (but expensive) design for interlaboratory method validation. (Collaborative trials typically cost €30 000 to conduct for one method.)

Proficiency testing schemes, however, differ from collaborative trials, in design and outcome, in a number of consequential ways.

- Often, only one test material (or a small number) is sent out in any one round, as opposed to a minimum of five in collaborative trials. It is, therefore, necessary to collect data from many rounds, over a period of perhaps several years, to provide sufficient information for validation purposes. (It is important to remember in this context that, strictly speaking, we do not validate "a method" as an isolated entity. What we validate is a method applied to specific analytes and defined ranges of test materials and analyte concentrations. Hence, not all rounds in a series may be eligible for use in the validation exercise.)

- Proficiency testing schemes rarely call for the reporting of duplicate results, so that estimates of repeatability standard deviation are not available from proficiency test results. (This is no great loss—it is simple for laboratories to estimate their own repeatability standard deviations.)
- In proficiency testing schemes, there is no guarantee that the same laboratories will participate in the scheme in different rounds.
- In a collaborative trial, the participants are selected on the basis of probable competence. In proficiency tests, universal competence is not a sensible assumption.

With due regard to these differences, the results of proficiency tests, restricted to participants using a closely defined method protocol, can reasonably be used to estimate reproducibility standard deviation for the method [2]. Robust estimation methods in combination with expert judgement are clearly called for to achieve the desired outcome. If two or more closely defined methods are used by sufficient numbers of participants, it is possible to assess any bias between the methods over an extended concentration range [3,4] by functional relationship estimation [5,6].

REFERENCES

1. W. Horwitz (Ed.). "Protocol for the design, conduct and interpretation of method performance studies", *Pure Appl. Chem.* **67**, 331–343 (1995).
2. Paper CX/MAS 02/12 of the Codex Committee on Methods of Analysis and Sampling. Validation of Methods Through the Use of Results from Proficiency Testing Schemes, Twenty-fourth Session, Budapest, Hungary, 18–22 November 2002, FAO, Rome.
3. P. J. Lowthian, M. Thompson, R. Wood. "The use of proficiency tests to assess the comparative performance of analytical methods: The determination of fat in foodstuffs", *Analyst* **121**, 977–982 (1996).
4. M. Thompson, L. Owen, K. Wilkinson, R. Wood, A. Damant. "A comparison of the Kjeldahl and Dumas methods for the determination of protein in foods, using data from a proficiency test", *Analyst* **127**, 1666–1668 (2002).
5. B. D. Ripley and M. Thompson. "Regression techniques for the detection of analytical bias", *Analyst* **112**, 377–383 (1987).
6. Analytical Methods Committee. "Fitting a linear functional relationship to data with error on both variables", AMC Technical Brief No 10. <www.rsc.org/amc/>.

APPENDIX 6: HOW PARTICIPANTS SHOULD RESPOND TO THE RESULTS OF PROFICIENCY TESTS

A6.1 General introduction

Taking part in a proficiency testing scheme is futile unless the participant makes full use of the results of each round, but avoids any misinterpretation. Proficiency testing is primarily a self-help tool that enables participants to detect unexpected sources of error in their results. However, it is not designed to be diagnostic. Consequently, it only helps a participant who is already using validated methods and has an IQC system in routine operation. Under those circumstances, an unexpected poor result in a proficiency test points to inadequacy in either the method validation or the IQC or, most likely, in both. (An adequate IQC system would normally flag up a problem with the analysis well before the proficiency test score was available. There is a demonstrable connection between a participant's performance in a proficiency test and the efficacy of the IQC system in use [1].)

Avoiding misunderstanding is especially important where the use of proficiency test scores goes beyond the purely scientific, and is used for example for accreditation or in a laboratory's promotional literature. The participant must take account of the statistical nature of z-scores in their interpretation.

The following guidelines may help participants with proper interpretation and use of *z*-scores. They are reproduced more-or-less intact from an AMC Technical Brief, with consent from the Royal Society of Chemistry [2].

A6.2 Proficiency testing and accreditation

Proficiency testing is so effective in detecting unexpected problems in analytical work that participation in a scheme (where one is available) is usually regarded as a prerequisite to accreditation for analytical work. Accreditation assessors will expect to see a documented system of appropriate responses to any results that show insufficient accuracy.

Such a system should include the following features:

- the definition of appropriate criteria for instigating investigative and/or remedial actions;
- the definition of the investigative and remedial procedures to be used and a scheme for their deployment;
- the recording of test results and conclusions accumulated during such investigations; and
- the recording of subsequent results showing that any remedial activities have been effective.

This section provides advice to enable analytical chemists to meet these needs and demonstrate that the needs have been met.

A6.3 Procedures and documentation

Participants should put in place a documented procedure for investigating and dealing with poor *z*-scores. Best practice for a participant will depend on exactly how the proficiency testing scheme is organized. The system could take the form of a flow chart or decision tree, based on the considerations discussed below and the participant's particular needs. However, the scope for the exercise of professional judgement should be included explicitly in the procedure.

Chemical proficiency testing schemes usually set a criterion for fitness-for-purpose that is broadly applicable over the relevant fields of application. Even if such a "fitness-for-purpose" criterion is set, it may or may not be appropriate for an individual participant's work for a particular customer. This factor needs to be considered when a participant sets up a formal system of response to the scores obtained in each round of a scheme. The main possibilities are covered below.

A6.4 Effect of scoring criterion

A6.4.1 The proficiency testing scheme uses a fitness-for-purpose criterion

The simplest possibility occurs when the scheme provides a criterion of fitness-for-purpose σ_p as a standard uncertainty and uses it to calculate *z*-scores. In this case it is important to realize that σ_p is determined in advance by the scheme organizers to describe their notion of fitness-for-purpose: It does not depend at all on the results obtained by the participants. The value of σ_p is determined so that it can be treated like a standard deviation. So if results are unbiased and distributed normally, and a participant's run-to-run standard deviation σ is equal to σ_p, then the *z*-scores will be distributed as $z \sim N(0,1)$, that is, on average about 1 in 20 of the *z*-scores fall outside the range ±2 and only about 3 in 1000 fall outside ±3.

Few (if any) laboratories fulfil these requirements exactly, however. For unbiased results, if a participant's run-to-run standard deviation σ is less than σ_p, then fewer points than specified above fall outside the respective limits. If $\sigma > \sigma_p$, then a greater proportion fall outside the limits. In practice, most participants operate under the condition $\sigma < \sigma_p$, but the results produced also include a bias of greater or smaller extent. Such biases often comprise the major part of the total error in a result, and they always serve to increase the proportion of results falling outside the limits. For example, in a laboratory where $\sigma = \sigma_p$, a bias of magnitude equal to σ_p will increase the proportion of results falling outside the $\pm 3\sigma_p$ limits by a factor of about eight.

Given these outcomes, it is clearly useful to record and interpret *z*-scores for a particular type of analysis in the form of a Shewhart [3] or other control chart (see also Appendix 4). If a participant's performance were consistently fit for purpose, a *z*-score outside the range ±3 would occur very rarely. If it did occur, it would be more reasonable to suppose that the analytical system produced a serious bias than a very unusual random error. The occurrence would demonstrate that the laboratory needed to take some kind of remedial action to eliminate the problem. Two successive *z*-scores falling between 2 and 3 (or between –2 and –3) could be interpreted in the same way. In fact, all of the normal rules for interpreting the Shewhart chart (e.g., the Westgard rules [3]) could be employed.

In addition to this use of the Shewhart chart, it may be worth testing *z*-scores for evidence of long-term bias as well, by using a cusum chart or a *J*-chart (Appendix 4). These bias tests are not strictly necessary: If a participant's *z*-scores nearly always fulfil the requirements of the fitness-for-purpose criterion, a small bias may not important. However, as we saw above, any degree of bias will tend to increase the proportion of results falling outside the action limits and may, therefore, be worth eliminating. A participant who decides to ignore the bias aspect should say so in the specification of investigatory actions. In other words, the participant should make it clear to the accreditation assessors that the decision to ignore bias is both deliberate and well founded rather than inadvertent.

A6.4.2 The proficiency testing scheme does not use an appropriate criterion of fitness-for-purpose

Some proficiency testing schemes do not operate on a "fitness-for-purpose" basis. The scheme provider calculates a score from the participants' results alone (i.e., with no external reference to actual requirements). More often, a participant may find that the fitness-for-purpose criterion used by the scheme provider is inappropriate for certain classes of work undertaken by the laboratory. Participants in such schemes may need to calculate their own scores based on fitness-for-purpose. That can be accomplished in a straightforward manner by the methods outlined below.

The participant should agree with the customer a specific fitness-for-purpose criterion in the form of a standard uncertainty σ_{ffp}, and use that to calculate the modified *z*-score given by

$$z_{\text{L}} = (x - x_{\text{a}})/\sigma_{\text{ffp}}$$

to replace the conventional *z*-score (see Section 3.5.4). The assigned value x_a should be obtained from the scheme itself. If there were several customers with different accuracy requirements, there could be several valid scores derived from any one result. These scores could be handled in exactly the manner recommended above for *z*-scores, that is, with the usual types of control chart. As the assigned value of the analyte is unknown to the participant at the time of analysis, a fitness-for-purpose criterion usually has to be specified as a function of *c*, the analyte concentration, as shown in Section 3.5.

A6.5 How to investigate a poor *z*-score

The investigation of a poor *z*-score is intimately connected with IQC [3]. In usual circumstances, a proficiency testing participant finds out about a poor *z*-score days or weeks after the run of analysis has taken place. In routine analysis, however, any substantive problem affecting the whole run should have been detected promptly by the IQC procedures. The cause of the problem would have been corrected immediately. The run containing the proficiency testing material would then have been reanalyzed, and a presumably more accurate result submitted to the proficiency testing scheme. So an *unexpectedly* poor *z*-score shows either that (a) the IQC system is inadequate, or (b) the proficiency testing material, alone of the test materials in the analytical run, was affected by a problem. Participants should consider both of these possibilities.

A6.5.1 Failings in IQC systems

A common failing of IQC is that the IQC material is poorly matched to the typical test material. An IQC material should be as far as possible representative of a typical test material, in respect of matrix, compartmentation, speciation, and concentration of the analyte. Only then can the behavior of the IQC material be a useful guide to that of the whole run. If the test materials vary greatly in any of these respects, use of more than one IQC material is beneficial. For instance, if the concentration of the analyte varies considerably among the test materials (say, over two orders of magnitude) two different IQC materials should be considered, with concentrations toward the extremes of the usual range. It is especially important to avoid using a simple standard solution of the analyte as an IQC surrogate for a test material with a complex matrix.

Another problem can arise if the IQC system addresses only between-run precision and neglects bias in the mean result. Such bias can result in a problem whether or not the IQC material is matrix-matched with the usual type of test material (and, by implication, with the proficiency testing material). It is, therefore, important to compare the mean result with the best possible estimate of the true value for the IQC material. Obtaining such an estimate requires traceability to outside the parent laboratory. External traceability could be obtained, for instance, by reference to CRMs of comparable matrix, or by subjecting the candidate IQC material to an interlaboratory study of some kind.

A6.5.2 A problem with the proficiency testing material

If the participant is satisfied that the IQC system is demonstrably unbiased, a problem unique to the proficiency testing material result must be suspected. The poor result could be the outcome of a mistake related to the handling of the proficiency testing material (e.g.,, an incorrect weight or volume recorded). That should be easily checked. Alternatively, an unexpected form of bias (such as a previously unobserved interference effect or unusually low recovery) might have uniquely affected the proficiency testing material or the measurement process. A valid conclusion at this stage might be that the proficiency testing material is sufficiently different from the typical test material to make the *z*-score inapplicable to the analytical task being undertaken.

A6.5.3 Diagnostic tests

A poor *z*-score is indicative of a problem, but is not diagnostic, so a participant usually requires further information to determine the origin of a poor result. As a first stage, a participant should re-examine the records for the run of analysis containing the proficiency testing material. The following features should be sought:

- systematic or sporadic mistakes in calculations;
- incorrect weights or volumes used;
- out-of-control indications from routine IQC charts;
- unusually high blanks; and
- poor recoveries, etc.

If these actions yield no insight, then further measurements are needed.

The obvious action is to reanalyze the proficiency testing material in question in the next routine run of analysis. If the problem disappears (i.e., the new result gives rise to an acceptable *z*-score), the participant may have to attribute the original problem to a sporadic event of unknown cause. If the poor result persists, a more extensive investigation is called for. That could be effected by the analysis of a run containing proficiency testing materials from previous rounds of the scheme and/or appropriate CRMs if they are available.

If the poor result is still obtained for the proficiency testing material under investigation, but is absent from the result for the other proficiency testing materials and CRMs, then it is likely to result from a unique property of the material, possibly an unexpected interference or matrix effect. Such a finding may call for more extensive studies to identify the cause of the interference. In addition, the participant may need to modify the routine analytical procedure to accommodate the presence of the interferent in future test materials. (However, it may know that routine test materials would never contain the interferent, and decide that the unfavorable *z*-score was inapplicable to the particular laboratory.)

If the problem is general among the results of the old proficiency testing materials and the CRMs, there is a probably a defect in the analytical procedure and a corresponding defect in the IQC system. Both of these would demand attention.

A6.5.4 Extra information from multi-analyte results

Some proficiency tests involve methods, such as atomic emission spectrophotometry, that can simultaneously determine of a number of analytes from a single test portion and a single chemical treatment. (Chromatographic methods that determine a number of analytes in quick succession can also be regarded as "simultaneous" in the present discussion.) Additional information that is diagnostic can sometimes be recovered from multianalyte results from a proficiency testing material. If all or most of the analytes have unsatisfactory results and are affected roughly to the same degree, the fault must lie in an action that affects the whole procedure, such as a mistake in the weighing of the test portion or in adding an internal standard. If only one analyte is adversely affected, the problem must lie in the calibration for that analyte or in a unique aspect of the chemistry of that analyte. If a substantial subset of the analytes is affected, the same factors apply. For instance, in elemental analysis of rock, if a group of elements give low results, it might be productive to see whether the effect could be traced to the incomplete dissolution of one of the mineral phases comprising the rock in which those elements are concentrated. Alternatively, there might be a spectrochemical change brought about by variation in the operation of the nebulizer system or the plasma itself that affects some elements rather than others.

A6.5.4 A suspected biased assigned value

Most proficiency testing schemes use a participant consensus as the assigned value. There is seldom a practicable alternative. However, the use of the consensus raises the possibility that there is, among a group of laboratories mainly using a biased analytical method, a small minority of participants that use a bias-free method. This minority subset can produce results that deviate from the consensus and generate "unacceptable" *z*-scores. In practice, such an occurrence is unusual but not unknown, particularly when new analytes or test materials are being subjected to proficiency testing. For instance, the majority of participants might use a method that is prone to an unrecognized interference, while the minority have detected the interference and developed a method that overcomes it.

Often the problem is immediately apparent to the participants affected, because they have used a method that is based on a deeper understanding of the chemical procedures than the one used by the majority of the participants. But the problem is not visible to other participants or the scheme provider. If a participant suspects that they are in this position, the correct course of action, having passed through the steps outlined above, is to send to the proficiency test provider details of the evidence accumulated that the assigned value is defective. The provider will normally have access to records of the methods used by the other participants and may be in a position to substantiate the complaint immediately. Alternatively, the provider may set in action a longer-term investigation into the problem, which should resolve the discrepancy in due course.

REFERENCES

1. M. Thompson and P. J. Lowthian. "Effectiveness of analytical quality control is related to the subsequent performance of laboratories in proficiency tests", *Analyst* **118**, 1495–1500 (1993).
2. Analytical Methods Committee. "Understanding and acting on scores obtained in proficiency testing schemes", AMC Technical Brief No 11. <www.rsc.org/amc/>.
3. M. Thompson and R. Wood. "Harmonised guidelines for internal quality control in analytical chemistry laboratories", *Pure Appl. Chem.* **67**, 649–666 (1995).

APPENDIX 7: GUIDE TO PROFICIENCY TESTING FOR END-USERS OF DATA

These questions and answers are based on misunderstandings reported by end-users of analytical data. The interpretation of proficiency test results in analytical chemistry should be conducted with the collaboration of an analytical chemist.

What is proficiency testing?
Proficiency testing comprises an interlaboratory system for the regular testing of the accuracy that the participant laboratories can achieve. In its usual form, the organizers of the scheme distribute portions of a homogeneous material to each the participants, who analyze the material under typical conditions and report the result to the organizers. The organizers compile the results and inform the participants of the outcome, usually in the form of a score relating to the accuracy of the result.

What is the difference between proficiency testing and accreditation?
Accreditation agencies require analytical laboratories to participate in an appropriate proficiency testing scheme where one is available, and demonstrate a system for handling the outcome. This is only one of many requirements of accreditation.

What kinds of materials are distributed?
The materials distributed are as close as possible to the materials being regularly analyzed, so that the results of the scheme represent the capability of the laboratories working under routine conditions.

What is proficiency testing for?
The primary purpose of proficiency testing is to help laboratories detect and cure any unacceptably large inaccuracy in their reported results. In other words, it is designed as a self-help system to tell the participants whether they need to modify their procedures. Proficiency tests are not ideally designed for any other purpose, although their results, with due regard to their limitations, can be used and combined with other information for certain other purposes.

Why are there inaccuracies in analytical results?
All measurement gives rise to inaccuracies, technically known as “errors” in the measurement community. (The word “error” here does not imply that a mistake has been made, merely that the outcome of the measurement process varies.) Errors arise because of unavoidable variation in the physical or chemical procedure employed to make the measurement. Measurements of chemical concentration require far more complicated procedures than typical physical measurements such as length or time. It is straightforward to measure a length to an accuracy of one part in a million, but chemical measurements can seldom be made with an accuracy of better than one part in a hundred. Mostly, the accuracy is not as good as that, especially if concentrations are very low, for instance, as when pesticide residues are being determined in foodstuffs.

Is the available accuracy good enough?
That depends on the application. Some analyses have to be extremely accurate. For example, in determining the commercial value of a consignment of scrap gold, the gold content has to be determined with the smallest possible error (less than one part in a thousand) because a small error could equate to many thousands of euros. In other applications, for example, in determining the concentration of copper in soil, an accuracy of one part in ten probably suffices–it doesn’t matter whether the true value is 20 or 22 ppm if the only decision is whether the level is above or below 200 ppm. Cost comes into consideration as well. As a rule of thumb, to improve the accuracy of a measurement by a factor of two decreases the chance of an incorrect (i.e., expensive) decision, but increases the cost of analysis by a factor of four. These considerations are known as “fitness-for-purpose”.

How do proficiency testing schemes evaluate the accuracy of individual laboratories?
Most schemes convert the participant's result into a "*z*-score". This score reflects two separate features, (a) the actual accuracy achieved (i.e., the difference between the participant's result and the accepted true value), and (b) the scheme organizer's judgement of what degree of accuracy is fit-for-purpose.

How should *z*-scores be interpreted?
z-Scores must be interpreted on a statistical (probabilistic) basis, and this requires expert knowledge. However, the following simple rules apply:

- A score of zero implies a perfect result. This will happen quite rarely even in perfectly competent laboratories.
- Laboratories complying with the proficiency testing scheme's fitness-for-purpose criterion will commonly produce scores falling between –2 and 2. They might expect to produce a value somewhat outside this range occasionally, roughly about 1 time in 20, so an isolated event of this kind is not of great moment. The sign (i.e., + or –) of the score indicates a negative or positive error, respectively
- A score outside the range from –3 to 3 would be very unusual for a laboratory operating under the given fitness-for-purpose criterion, and is taken to indicate that the accuracy requirement has not been met (at least on that occasion). The cause of the event should be investigated and remedied.

What mistakes are commonly made in using *z*-scores?
It is important not to over-interpret *z*-scores. This could happen in a number of ways, such as the following:

- Comparing *z*-scores between rounds or between laboratories has to be done with great caution. A single laboratory operating consistently in line with the fitness-for-purpose criterion would typically produce *z*-scores in successive rounds covering the range –2 to +2: The following set [0.6, –0.8, 0.3, 1.7, 0.7, –0.1] would be typical. The small ups and downs between the scores do not indicate a change in performance—they arise by chance. So 1.7 is not "worse" than 0.3, and it does not indicate deterioration in performance.
- Because of this "natural variation", it is not valid to make a "league table" of laboratories based on their *z*-scores in a round. It is not valid to claim that a laboratory scoring 0.3 in a single round is better than another scoring 1.7.
- Judgements based on average *z*-scores again require caution. Averages of *z*-scores obtained on a number of different analytes should not be used; they may well hide the fact that one of the analytes consistently gives a poor *z*-score. Averages of scores from the same analyte over several rounds may be more useful, but still need expert interpretation.

What are the limitations of proficiency testing?

- Proficiency testing has to be carried out within the context of a complete system for appropriate quality in each laboratory. It cannot be used as a substitute for routine QC. It is not, in isolation, a sufficient means of validating analytical methods, nor of training individual analysts.
- Proficiency testing provides a participant laboratory only with an indication of problems if they are present. It does not provide any diagnostics to help solve the problem.
- Success in a proficiency test for one analyte does not indicate that a laboratory is equally competent in determining an unrelated analyte.

List of symbols

C	Cochran's test statistic for duplicate data ($C = D_{\mathrm{max}}/S_{\mathrm{DD}}$)
c	critical value in a test for sufficient homogeneity ($c = F_1\sigma^2_{\mathrm{all}} + F_2 s^2_{\mathrm{an}}$)
c_{L}	arbitrarily determined limit for reporting
D_i	difference of i^{th} pair of duplicates in a homogeneity test
D_{max}	largest difference of pairs of duplicates
f	constant used in setting standard deviation for proficiency testing
F	ratio of sample variances used in the F-test for equality of variance
F_1, F_2	critical values for homogeneity testing (Appendix 1)
H	statistical hypothesis
H_0	statistical null hypothesis
h	bandwidth in a kernel density
J-score	score based on number of successive results on either side of zero line
l	multiplier for use in deciding upper limits on acceptable uncertainty for assigned value
m	number of distribution units in a homogeneity test
$N(\mu,\sigma^2)$	normal distribution with population mean μ and population variance σ^2
n	number of results
s_{an}	experimental estimate of analytical standard deviation
s_{sam}	experimental estimate of sampling standard deviation
S_{DD}	sum of squared differences in homogeneity test based on duplicate analysis
S_{ZZ}	sum of squared z-scores $S_{ZZ} = \sum_i z_i^2$
$S_{\mathrm{Z,rs}}$	rescaled sum of z-scores $= \sum_i z_i/\sqrt{n}$
t	student's t-statistic
$u(x)$	standard uncertainty in x
V_{S}	variance of sums in a homogeneity test $= \Sigma(S_i - \bar{S})^2/(m - 1)$
x	participant's result
x_{a}	assigned value
x_{max}	maximum allowable analyte concentration
x_{true}	true value of measured quantity
z	z-score $z = (x - x_{\mathrm{a}})/\sigma_{\mathrm{p}}$
z'	modified z-score including assigned value uncertainty, $z' = \dfrac{x - x_{\mathrm{a}}}{\sqrt{u^2(x_{\mathrm{a}}) + \sigma^2_{\mathrm{p}}}}$
z_{L}	modified z-score incorporating laboratory-specific performance criterion $z_{\mathrm{L}} = (x - x_{\mathrm{a}})/\ \sigma_{\mathrm{ffp}}$
μ_{cont}	population mean of results for material stored under stable control conditions
μ_{expt}	population mean of results for material under stability test experimental conditions
$\hat{\mu}_{\mathrm{rob}}$	robust mean
σ	population standard deviation, in general
σ_{all}	allowed standard deviation
σ_{ffp}	fitness-for-purpose standard deviation
σ_{p}	standard deviation for proficiency testing
$\hat{\sigma}_{\mathrm{rob}}$	robust standard deviation
σ_{sam}	(true) sampling standard deviation, i.e., contribution of sample-to-sample variation to observed dispersion of observation
χ^2_n	chi-square distribution with n degrees of freedom
ζ	zeta score, $\zeta = (x - x_{\mathrm{a}})/\sqrt{u^2\ (x) + u^2(x_{\mathrm{a}})}$