amc technical brief

Analytical Methods Committee No.11 Dec 2002

Understanding and acting on scores obtained in proficiency testing schemes

Proficiency testing (PT) is so effective in detecting unexpected problems in analytical work that participation in a scheme (where one is available) is regarded as a prerequisite to accreditation. Moreover, as well as evidence that a laboratory is participating a in PT scheme, 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 investigatory and/or remedial actions;
- the definition of the investigatory and remedial procedures to be used and a scheme for their deployment;
- the recording the test results and conclusions accumulated during such investigations; and
- the recording of subsequent results showing that any remedial activities have been effective.

This technical brief provides the background to enable analytical chemists to meet these needs and demonstrate that the needs have been met. However, because of variations in practice among PT schemes, the statistical basis of proficiency testing is not quite as simple as it is usually presented. It is therefore important for everybody concerned to understand exactly how a particular scheme is organised. The main possibilities are covered below. One of the key issues is whether the PT scheme is using a fitness-for-purpose criterion that is appropriate for the individual participant's requirements.

Fitness for purpose (FFP)

The primary purpose of proficiency testing[1-3] in chemical analysis is to provide a means by which participant laboratories can regularly check that their results are fit for purpose. Fitness for purpose implies that the uncertainty is sufficiently small that correct decisions can be based on analytical results without undue expenditure on the measurement.[4] The level of uncertainty that comprises fitness for purpose is therefore a matter that should be agreed between the laboratory and the customer before any analysis is undertaken. Chemical proficiency testing schemes usually set a standard for fitness for purpose that is broadly applicable over the relevant fields of application. However, that standard may or may not be appropriate for an individual participant's work for a particular customer.

These factors need to be considered when a participant sets up a formal system of response to the scores obtained in each round of a scheme. We therefore need to consider three commonly encountered situations:

- the PT scheme uses an appropriate FFP criterion;
- the scheme does not use a FFP criterion;
- the scheme uses an inappropriate FFP criterion.

The PT scheme uses an appropriate FFP criterion

The simplest possibility occurs when the scheme provides a criterion of fitness for purpose s_p as a standard uncertainty and uses it to calculate z-scores from the equation

$$z = (x - X) / s_p,$$

where x is the participant's result and X is the assigned value. In this case it is important to realise that the target value s_p is determined in advance by the scheme organisers to describe their notion of fitness for purpose: it does not depend at all on the results obtained by the participants. The value of s_p is determined so that it can be treated like a standard deviation. So if your result is unbiased and distributed normally, and your run-to-run standard deviation s is equal to s_p, then your z-scores will be $z \sim N(0,1)$, *i.e.*, taken at random from a normal distribution with zero mean and unit variance. On average, about 1 in 20 of such 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 s is less than s_p, then fewer points than specified above fall outside the respective limits. If $s > s_p$, then a greater proportion would fall outside the limits. In reality, most participants operate under the condition $s < s_p$, but their results 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 $s = s_p$, a bias of magnitude equal to s_p will increase the proportion of results falling outside the $\pm 3s_p$ 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 control chart[5] (see below).

The PT scheme does not use a 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). In such a scheme, you might find a z-score calculated by using a standard deviation estimated from the participants' results (with appropriate treatment of outliers) as the value of s_p. That strategy ensures that about 95% of participants always get an apparently "satisfactory" score (*i.e.*, in the range ±2), regardless of whether the accuracy is appropriate. That may be comforting for the participants (and, indeed, for the scheme provider) but it says nothing about whether the results are fit for purpose. Alternatively a "q-score" can be calculated, simply a relative error given by $q = (x - X)/X$. Again, this says nothing about fitness for purpose.

If your PT scheme operates on this kind of basis, you need to calculate your own score based on fitness for purpose. That can be accomplished in a straightforward manner by the methods outlined in the next section.

The PT scheme uses an inappropriate criterion

More often than a PT scheme having no fitness-for-purpose criterion, 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. If that happens, the participant should agree a specific fitness-for-purpose criterion s_f with the customer, and use that to calculate the 'zeta-score', given by

$$z = (x - X)/s_f,$$

to replace the conventional z-score.[6] As before, x is the participant's result and X is the scheme's assigned value. The criterion s_f 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 zeta-scores derived from any one result. These zeta-scores could be handled in exactly the manner recommended above for z-scores, that is, with the usual types of control chart.

Concentration dependency

As the concentration of the analyte is unknown to the participant at the time of analysis, a fitness-for-purpose criterion usually has to be expressed as concentration-dependent. You simply need to specify the fitness-for-purpose criterion as a function of *c*, the analyte concentration. For example, you might need a constant relative standard deviation, giving

$$s_f = Ac,$$

where A is an agreed constant. We could find the appropriate value of s_f by using the assigned value given by the scheme as the concentration, *i.e.*, $c = X$.

A more elaborate function might take note of the fact that there is often a lower limit of concentration c_L below which a less stringent uncertainty requirement is satisfactory. In that instance a relationship of the form

$$s_f = c_L/B + Ac$$

might be more satisfactory, where *B* is another constant. This would ensure that s_f could not fall below c_L/B, however low the actual concentration of analyte. Another possibility would be to use the Horwitz equation[7]

$$s_f = 0.02c^{0.8495},$$

or an analogous equation, as the fitness function. (Note that in the Horwitz equation, both c and s_f are in unit-free mass fractions.)

Control Charts

If a laboratory'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 had produced a serious bias than a very unusual random error. Such an 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 (for example the Westgard Rules[5]) could be employed. In practice, a laboratory may wish to set action limits at a point between 2 and 3, to correspond with an intermediate chosen level of probability.

In addition to this use of the Shewhart chart, it is often worth testing z-scores for evidence of long term bias as well, for instance by using a cusum chart[8] or a J-chart.[8] However, 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 that the decision to ignore bias is deliberate and well-founded rather than inadvertent.

How to respond to a z-score requiring action

The investigation of a poor z-score is intimately connected with internal quality control (IQC).[5] In usual circumstances,

a PT participant finds out about a poor z-score days or weeks after the run of analysis took place. In routine analysis, however, any extensive problem affecting the whole run should have been detected promptly by the internal quality control procedures. The cause of the problem would have been corrected immediately. The run containing the PT material would then have been reanalysed, and a presumably more accurate result submitted to the PT scheme. So an *unexpectedly* poor z-score shows either that (a) the IQC system is inadequate, or (b) the PT material, alone of the test materials in the analytical run, was affected by a problem. Participants should consider both of these possibilities.

Failings in internal quality control (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, compartment, speciation and concentration of the analyte. Only then can the behaviour 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 within the defined class, 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 roughly at the quartiles 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 PT 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 a 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.

An unusual PT material

If the participant is satisfied that the IQC system is demonstrably unbiased, the problem with the PT material result might be unique to that particular analytical result. The poor result could be the outcome of a mistake related to the handling of the PT material (for example, an incorrect weight or volume recorded). That could be quickly checked. Alternatively, an unexpected form of bias (such as a previously unobserved interference effect or unusually low recovery) might have uniquely affected the PT material or the measurement process. A tentative conclusion to be drawn in the latter case might be that the PT material is sufficiently different from the typical test material to make the z-score inapplicable to the analytical task being undertaken. The alternative is that the analytical method and the IQC system need modification.

Diagnostic tests

A poor z-score is indicative of a problem, but is not diagnostic, so you usually require further information to determine the origin of a poor result. As a first stage you 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 your routine IQC charts;
- unusually high blanks;
- poor recoveries, etc.

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

The obvious action is to reanalyse the PT 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), you 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. You could effect that by the analysis of a run containing PT materials from previous rounds of the scheme and/or appropriate CRMs if they are available.

If the poor result is still obtained for the PT material under investigation, but is absent from the result for the other PT 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, you may need to modify the routine analytical procedure to accommodate the presence of the interferent in future test materials. (However, you may know that your own test materials would never contain the interferent, and decide that the unfavourable z-score was inapplicable to your analytical system.)

If the problem is general among the results of the old PT 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.

Extra information from multi-analyte results

Some proficiency tests involve methods, such as ICPAES, that can simultaneously determine 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 PT 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 the analysis of a rock by ICPAES, if a group of elements gives low results, it might be productive to see whether the effect could be traced to the incomplete dissolution of one of the mineral phases making up the rock in which those elements are concentrated. Alternatively, there might be a spectrochemical change brought about by variation in the operation of the nebuliser system or the plasma itself that affects some elements rather than others.

A biased assigned value

Ideally proficiency testing schemes should employ traceable assigned values. In practice, most proficiency testing schemes use a participant consensus as the assigned value, because there is seldom a practicable alternative. However, the use of the consensus raises the theoretical 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 produce results that deviate from the consensus and unfairly 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 unrecognised 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 immediately to substantiate the complaint. Alternatively, the provider may set in action a longer-term investigation into the problem. Hopefully that would resolve the discrepancy in due course. Such an event should not be regarded as a defect in proficiency testing but, in fact, one of its benefits - a problem that was not apparent to laboratories working in isolation has been discovered and rectified.

Conclusions

Participants should have a documented procedure for investigating and dealing with 'unsatisfactory' z-scores. This could, perhaps, take the form of a flow chart or decision tree, based on the considerations discussed above. Interpretation of results should take into account the participant laboratory's own fitness-for-purpose requirements. Inspection of IQC results and reanalysis of PT materials are recommended for supportive actions. However, we must recognise that no exactly-defined procedure can take account of every possible contingency Therefore scope for the exercise of professional judgement should be included explicitly in the procedure.

References

1 ISO Guide 43, *Proficiency Testing by Interlaboratory Comparisons*, ISO, Geneva, 1997.
2 M Thompson, R Wood, *Pure Appl. Chem.*, 1993, **65**, 2123.
3 R E Lawn, M Thompson, R F Walker, *Proficiency Testing in Analytical Chemistry*, Royal Society of Chemistry, Cambridge, 1997.
4 T Fearn, S A Fisher, M Thompson and S L R Ellison, *Analyst*, 2002, **127**, 818-824.
5 M Thompson and R Wood, *Pure Appl Chem*, 1995, **67**, 649-666.
6 AMC Technical Briefs, 2000, No 2, www.rsc.org/lap/rsccom/amc/amc_index.htm
7 W Horwitz and R Albert, *J. AOAC Int*., 1996, **79**, 589.
6 R J Howarth, *Analyst*, 1995, **120**, 1851-1873.

About the AMC: The Analytical Methods Committee handles matters that are of technical importance to the Analytical Division of the RSC and the analytical community in general. The aim of the AMC is "to participate in national and international efforts to establish a comprehensive framework for appropriate quality in chemical measurement". It achieves this aim through the activities of its expert subcommittees, which handle:

- the development, revision and promulgation of validated, standardised and official methods of analysis;
- the development and establishment of suitable performance criteria for methods and instruments;
- the use and development of appropriate statistical methods;
- the identification and promulgation of best analytical practice, including aspects relating to sampling, equipment, instrumentation and materials;
- the generation of validated compositional data of natural products for interpretative purposes.