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.
purpose
p

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

ttonsi0ered, wiIwould hds out abouhedse Psfical terial is poequate, o 0.0055 Tc 0.1145 7a Tj comorals in tdematste testunt intrueuld hav