GUIDE TO PROFICIENCY TESTING AUSTRALIA

2024

Revised February 2024

PROFICIENCY TESTING AUSTRALIA
PO Box 7507 Silverwater NSW 2128 AUSTRALIA

CONTENTS

1. Scope

The purpose of this document is to provide participants in Proficiency Testing Australia's (PTA) programs with an overview of how the various types of proficiency testing programs are conducted and an explanation of how laboratory performance is evaluated. The document does not attempt to cover each step in the proficiency testing process. These are covered in PTA's internal procedures which are in compliance with the requirements of ISO/IEC 17043[1].

The main body of this document contains general information about PTA's programs and is intended for all users of this document. The appendices contain: a glossary of terms (A); information on the evaluation procedures used for testing programs (B); and details of the evaluation of the results for calibration programs (C).

2. Introduction

The competence of laboratories is assessed by two complementary techniques. One technique is an on-site evaluation to the requirements of ISO/IEC 17025[3]. The other technique is by proficiency testing which involves the determination of laboratory performance by means of interlaboratory comparisons, whereby the laboratory undergoes practical tests, and their results are compared with those of other laboratories. The two techniques each have their own advantages which, when combined, give a high degree of confidence in the integrity and effectiveness of the assessment process. Although proficiency testing schemes may often also provide information for other purposes (e.g. method evaluation), PTA uses them specifically for the determination of laboratory performance.

PTA programs are divided into two different categories - testing interlaboratory comparisons, which involve concurrent testing of samples by two or more laboratories and calculation of consensus values from all participants' results, and calibration interlaboratory comparisons in which one test item is distributed sequentially among two or more participating laboratories and each laboratory's results are compared to reference values. A subset of interlaboratory comparisons are one-off practical tests (refer Section 5.8) and measurement audits (refer Section 6.10) where a well characterised test item is distributed to *one* laboratory and the results are compared to reference values.

Proficiency testing is carried out by PTA staff. Technical input for each program is provided by Technical Advisers. The programs are conducted using collaborators for the supply and characterisation of the samples and test items. All other activities are undertaken by PTA.

2.1 Confidentiality

All information supplied by a laboratory as part of a proficiency testing program is treated as confidential. There are, however, three exceptions. Information can be disclosed to third parties:

- with the express approval of the client(s);
- when PTA has an agreement with or requirement in writing from the Commonwealth or a State Government which requires the provision of information, and the relevant parties/clients have been informed in writing of such agreement or requirement;
- when PTA has any concerns about the conduct of any aspect of the proficiency testing process or in relation to any safety, medical or public health issues identified in the proficiency testing process.

PTA sample suppliers, distributers and Technical Advisers are required to sign confidentiality declarations at the commencement of each program round.

2.2 Funding

PTA charges a participation fee for each program. This fee varies from program to program and participants are notified accordingly, prior to a program's commencement.

3. References

1. ISO/IEC 17043:2010 *Conformity assessment: General requirements for proficiency testing*

2. ISO/IEC 17043:2023 *Conformity assessment: General requirements for the competence of proficiency testing providers*

3. ISO/IEC 17025:2017 *General requirements for the competence of testing and calibration laboratories*

4. ISO/IEC Guide 98-3:2008 *Uncertainty of measurement – Part 3: Guide to the expression of uncertainty in measurement (GUM)*

5. ISO 13528:2022 *Statistical methods for use in proficiency testing by interlaboratory comparisons*

4. Quality Management of Proficiency Testing Schemes

In accordance with best international practice, PTA maintains and documents a quality system for the conduct of its proficiency testing programs. This quality system complies with the requirements specified in ISO/IEC 17043:2010[1].

5. Testing Interlaboratory Comparisons

5.1 Introduction

PTA uses collaborators for the supply and homogeneity testing of samples. All other activities are undertaken by PTA and technical input is provided by program Technical Advisers.

In the majority of interlaboratory comparisons conducted by PTA, subdivided samples (taken from a bulk sample) are distributed to participating laboratories which test these concurrently. They then return results to PTA for analysis and consensus values are determined.

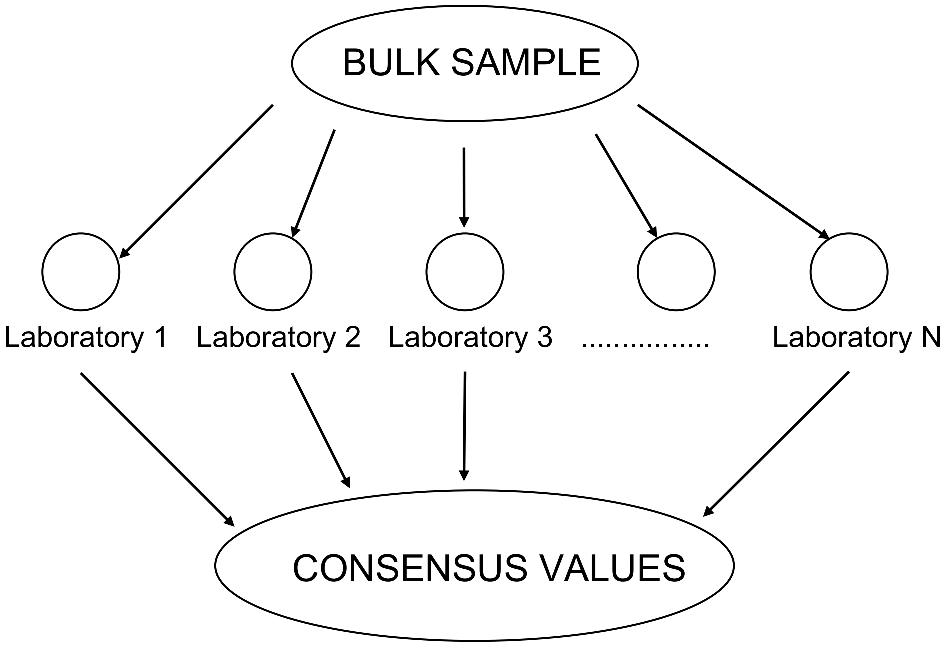

Figure 1: Typical Testing Interlaboratory Comparison

5.2 Working Group and Program Design

Once a program has been selected, a working group is formed. This group usually comprises one or more Technical Advisers, and the PTA staff member who will act as the Program Coordinator.

Technical Advisers provide input in the following areas:

- nomination of tests to be conducted, range of values to be included, test methods to be used and number/design of samples required;
- preparation of paperwork (instructions and results sheet) particularly with reference to reporting formats, number of decimal places to which results should be reported and correct units for reporting;
- identification and resolution of any difficulties expected in the preparation and maintenance of homogeneous proficiency test items, or in the provision of a stable assigned value for a proficiency test item;
- technical commentary in the final report and, in some cases, in answering questions from participants.

An appropriate statistical design is essential and therefore is established during the preliminary stages of the program (see Appendix B for further details).

5.3 Sample Supply and Preparation

The Program Coordinator is responsible for organising the supply and preparation of the samples. It is often the case that one of the Technical Advisers will also act as the program's sample supplier. In any case, the organisation preparing the test items is always one that is considered by PTA to have demonstrable competence to do so.

Sample preparation procedures are designed to ensure that the samples used are as homogeneous and stable as possible, while still being similar to samples routinely tested by laboratories. A number of samples are selected at random and tested, to ensure that they are sufficiently homogeneous for use in the proficiency program. Whenever possible, this is done prior to samples being distributed to participants. The results of this homogeneity testing are analysed statistically and may be included in the final report.

5.4 Packaging and Dispatch of Samples

The packaging and method of transport of the samples are considered carefully to ensure that they are adequate and able to protect the stability and characteristics of the samples. In some cases, samples are packaged and dispatched from the organisation supplying them, in other cases they are shipped to PTA for distribution. It is also ensured that certain restrictions on transport such as dangerous goods regulations or customs requirements are complied with.

5.5 Receipt of Results

Results from participating laboratories are required to be sent by email to the program coordinator. A 'due date' for return of results is set for each program, usually allowing laboratories two to three weeks to test the samples. If any results are outstanding after the due date, reminders are issued, however, as late results delay the data analysis, these may not be included. Laboratories are requested to submit all results on time.

5.6 Analysis of Data and Reporting of Results

Results are usually analysed together (with necessary distinctions made for method variation) to obtain consensus values for the entire group. The results received from participating laboratories are entered and analysed as soon as practicable so that the final report can be issued to participants within program timeframe.

The evaluation of the results is by calculation of robust z-scores, which are used to identify any outliers. Summary statistics and charts of the data are also produced, to assist with interpretation of the results. A detailed account of the procedures used to analyse results appears in Appendix B.

Participants are issued with an individual laboratory summary sheet which indicates which, if any, of their results were identified as outlier results. Where appropriate, it also includes other relevant comments (e.g. reporting logistics, method selection).

A final report is produced at the completion of a program and includes data on the distribution of results from all laboratories, together with an indication of each participant's performance. This report typically contains the following information:

(a) introduction;

(b) features of the program - number of participants, sample description, tests that were carried out;

(c) results from participants;

(d) statistical analysis, including graphical displays and data summaries (outlined in Appendix B);

(e) a table summarising the outlier† results;

(f) PTA and Technical Adviser's comments (on possible causes of outliers, variation between methods, overall performance etc.);

(g) sample preparation and homogeneity testing information; and

(h) a copy of the instructions to participants and results sheet.

Note: † *Outlier results are the results which are judged inconsistent with the consensus values (refer Appendix A for definition).*

The final program report is released by email to participants in the program.

5.7 Other Types of Testing Programs

PTA conducts some proficiency testing activities which do not exactly fit the model outlined in Section 5.1. These include known-value programs where samples with well-established reference values are distributed (e.g. slides for asbestos fibre counting).

Some of PTA's testing Interlaboratory comparisons may supply a certified reference material as the sample for testing. In some cases, the evaluation of results may be by En number (refer to Appendix C).

Some other PTA testing interlaboratory comparisons do not produce quantitative results - i.e. qualitative programs where the presence or absence of a particular parameter is to be determined (e.g. pathogens in food). By their nature, the results are also treated differently from the procedures outlined in Appendix B.

6. Calibration Interlaboratory Comparisons

6.1 Introduction

PTA uses collaborators for the supply and calibration of test items. All other activities are undertaken by PTA and technical input is provided by program Technical Advisers. Each calibration laboratory has its capability uniquely expressed both in terms of its ranges of measurements and the least measurement uncertainty (or best accuracy) applicable in each range. Because calibration laboratories are generally working to different levels of accuracy, it is not normally practicable to compare results on a group basis such as in interlaboratory *testing* programs. For calibration programs, we need to determine each individual laboratory's ability to achieve the level of accuracy for which they have nominated (their *least measurement uncertainties*).

The assigned (reference) values for a calibration program are not derived from a statistical analysis of the group's results. Instead, they are provided by a Reference Laboratory which must have a higher accuracy than that of the participating laboratories. For PTA interlaboratory comparisons, the Reference Laboratory is usually Australia's National Measurement Institute (NMI), which maintains Australia's primary standards of measurement.

6.2 Program Design

Once a program has been selected, a working group is formed. This group usually comprises one or more Technical Advisers and a PTA staff member who will act as the Program Coordinator. The group decides on the measurements to be conducted, how often the test item will need to be recalibrated and the range of values to be measured. They also formulate instructions and results sheets. PTA programs are designed so that it will normally take no more than eight hours for each participant to complete the measurements.

6.3 Test Item Selection

Because there can often be a substantial difference in the nominated measurement uncertainties of the participating laboratories, the test item must be carefully chosen. A test item with high resolution, good repeatability, good stability and an error that is large enough to be a meaningful test for all participants should be selected.

In some intercomparisons (especially international ones), the purpose may not only be to determine how well laboratories can measure specific points but also to highlight differences in methodology and interpretation.

6.4 Evaluation of Performance

As stated in Section 6.1, calibration laboratories are generally working to different levels of accuracy. Consequently, their performance is *not* judged by comparing their results with those of the other laboratories in an interlaboratory comparison. Instead, their results are compared only to the Reference Laboratory's results and their ability to achieve the accuracy for which they have nominated is evaluated by calculating the E_n number. For further details please refer to Appendix C.

6.5 Reference Values

Australia’s National Measurement Institute (NMI) provides most of the reference values for PTA’s Calibration interlaboratory comparisons. The majority of the participating laboratories’ reference equipment is also calibrated by NMI.

Test items with high resolution, good repeatability and good stability are selected. This is to ensure that these factors do not contribute significantly to the reference value uncertainty. Likewise, the Reference Laboratory has the capability to assign measurement uncertainties that are better than the participating laboratories.

6.6 Measurement Uncertainty (MU)

To be able to adequately compare laboratories they must report their uncertainties with the same confidence level. A confidence level of 95% is the most commonly used internationally. Laboratories should also use the same procedures to estimate their uncertainties as given in the ISO Guide[4].

Laboratories should not report uncertainties smaller than their nominated measurement uncertainty.

6.7 Reporting

An individual summary sheet is sent to laboratories to give them feedback on their performance. The summary sheet states the E_n values for each measurement based on the preliminary reference values and usually does not contain any technical commentary.

A *Final Report* is issued by email to participants, at the conclusion of the program. This typically contains more information than is provided in the summary sheet - including technical commentary and graphical displays.

6.8 Measurement Audits

The term *measurement audit* is used by PTA to describe a practical test whereby a well characterised and calibrated test item (or artefact) is sent to a single laboratory and the results are compared with a reference value.

Procedures are the same as for a normal interlaboratory comparison except that usually only a simple report is generated.

APPENDIX A

GLOSSARY OF TERMS

GLOSSARY OF TERMS

Further details about many of these terms may be found in either Appendix B (testing programs) or Appendix C (calibration programs). A number of these are also defined in ISO/IEC 17043[2].

assigned value value attributed to a particular property or characteristic of a proficiency testing item

consensus value value derived from a collection of results in an interlaboratory comparison

E_n number stands for error normalised; a quantitative measure of laboratory performance for calibration programs (see formula in Appendix C). The En can be useful for other types of proficiency testing

interlaboratory comparison design, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions

measurement uncertainty (MU) non-negative parameter characterising the dispersion of the quantity values being attributed to a measurand, based on the information used

outlier member of a set of values which is inconsistent with other members of that set

reference value an assigned value which is provided by a Reference Laboratory; quantity value used as a basis for comparison with values of quantities of the same kind

robust statistics statistical method insensitive to small departures from underlying assumptions surrounding an underlying probabilistic model

z-score (Z) a normalised value which assigns a “score” to the result(s), relative to the other numbers in the group

APPENDIX B

EVALUATION PROCEDURES FOR TESTING PROGRAMS

B.1 Introduction

This appendix outlines the procedures PTA uses to analyse the results of its proficiency testing programs. It is important to note that these procedures are applied only to *testing* programs, not *calibration* programs (which are covered in Appendix C). In testing programs, the evaluation of results is based on comparison to assigned values which are usually obtained from all participants' results (i.e. consensus values).

The statistical procedures described in this appendix have been chosen so that they can be applied to a wide range of testing programs and, whenever practicable, programs are designed so that these 'standard' procedures can be used to analyse the results. In some cases, however, a program is run where the 'standard' statistical analyses cannot be applied - in these cases other, more appropriate, statistical procedures may be used.

For all programs the statistical analysis is only one part of the evaluation of the results. If a result is identified as an outlier, this means that statistically it is significantly different from the others in the group, however, from the point of view of the specific science involved (e.g. chemistry), there may be nothing "wrong" with this result. This is why the assessment of the results is always a combination of the statistical analysis and input by Technical Advisers (who are experts in the field). In most cases the Technical Adviser's assessment matches the statistical assessment.

B.2 Statistical Design

In order to assess the testing performance of laboratories in a program, a robust statistical approach, using z-scores, is used. Z-scores give a measure of how far a result is from the assigned value and give a "score" to each result relative to the other results in the group. Section B.5 describes the method used by PTA for calculating z-scores.

For most testing programs, simple robust z-scores are calculated for each sample. Occasionally, the samples in a program may be paired and robust z-scores can be calculated for the sample pair. If paired samples are used, they may be identical ("blind duplicates") or slightly different (i.e. the properties to be tested are at different levels). The pairs of results which are subsequently obtained fall into two categories: uniform pairs, where the results are expected to be the same (i.e. the samples are identical or the same sample has been tested twice); and split pairs, where the results should be slightly different. The pairing of samples allows the assessment of both between-laboratories and within-laboratory variation in a program.

One of the main statistical considerations made during the planning of a program is that the analysis used is based on the assumption that the results will be approximately normally distributed. This means that the results roughly follow a normal distribution, which is the most common type of statistical distribution (see Figure 2).

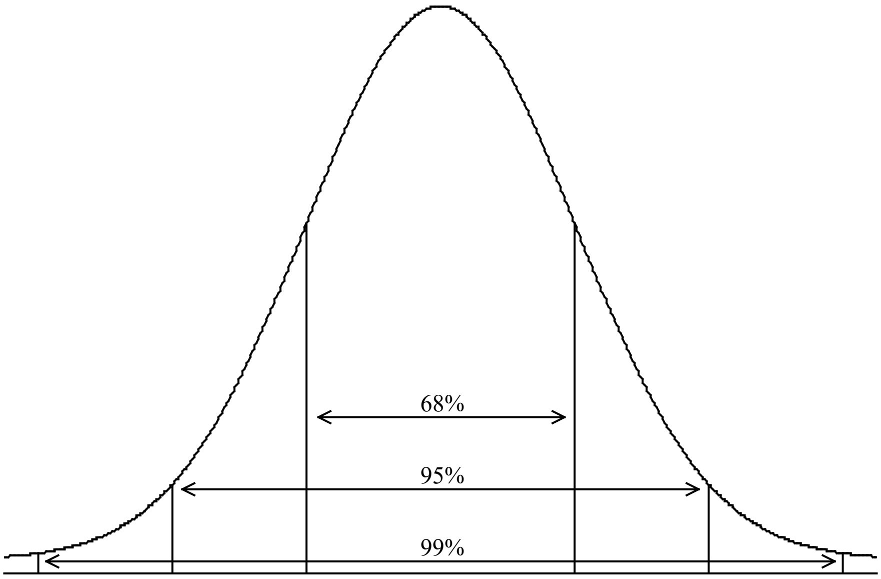

Figure 2: The Normal Distribution

The normal distribution is a “bell-shaped” curve, which is continuous and symmetric, and is defined such that about 68% of the values lie within one standard deviation of the mean, 95% are within two standard deviations and 99% are within three. To ensure that the results for a program are approximately normally distributed, the working group (in particular the Technical Adviser) considers carefully the results which might be obtained for the samples which are to be used.

For example, for the results to be continuous, careful consideration must be given to the units and number of decimal places requested - otherwise the data may contain a large number of repeated values. Another problem which should be avoided is when the properties to be tested are at very low levels - in this case the results are often not symmetric (i.e. skewed towards zero).

B.3 Data Preparation

Prior to commencing the statistical analysis, a number of steps are undertaken to ensure that the data collected is accurate and appropriate for analysis.

As the results are submitted to PTA, care is taken to ensure that all results are entered correctly. Once all of the results have been received (or the deadline for submission has passed), the entered results are carefully double-checked. It is during this checking phase that gross errors and potential problems with the data in general may be identified.

In some cases, the results are then transformed - for example, for microbiological count data the statistical analysis is usually carried out on the $\log_{10}$ of the results, rather than the raw counts. When all results have been entered and checked (and transformed if necessary) histograms of the data - which indicate the distribution of the results - are generated to check the assumption of normality.

These histograms are examined to see whether the results are continuous and symmetric. If this is not the case the statistical analysis may not be valid. In one case, two distinct groups of results are present on the histogram (i.e. a bi-modal distribution). This is most commonly due to two test methods giving different results, and in this case, it may be possible to separate the results for the two methods and then perform the statistical analysis on each group.

B.4 Summary Statistics

Once the data preparation is complete, summary statistics are calculated to describe the data. PTA uses robust statistics, which means that they are not influenced by the presence of outliers in the data set.

B.5 Robust Z-scores and Outliers

To statistically evaluate the participants' results, PTA uses z-scores based on robust summary statistics.

The calculated z-scores are tabulated in the report for a program, alongside the corresponding results and the results are assessed based on their z-scores. The interpretation of z-scores is as below:

$|Z| \leq 2.0$	indicates a "satisfactory" performance
$2.0 < |Z| < 3.0$	indicates a "questionable" performance
$|Z| \geq 3.0$	indicates an "unsatisfactory" performance

where |Z| denotes the absolute value of the z-score.

An outlier is defined as any result with an absolute z-score greater than or equal to three, i.e. $Z \geq 3.0$ or $Z \leq -3.0$. Outliers are identified in the tabulated results in a report by a marker (**§**) beside the z-score. When an outlier is identified the sign of the z-score indicates whether the result is too high (positive z-score) or too low (negative z-score). Laboratories that obtain outliers or questionable results in a program are encouraged to review their results.

In some circumstances, if the spread of results is too large or too small in the opinion of the Technical Adviser, a target coefficient of variation (CV) is used to calculate z-scores.

The actual value used as the target CV to calculate such z-scores is chosen in consultation with the Technical Adviser and usually takes into account historical data (most likely obtained from previous rounds of the program, or similar interlaboratory testing programs).

When pairs of results have been obtained in a program, two z-scores may be calculated - a between-laboratories z-score and a within-laboratory z-score. These are based on the sum and difference of the pair of results, respectively.

B.6 Graphical Displays

A number of graphical displays of the data are included in the report for a program. The two most commonly used graphs are the ordered z-score bar-chart and the Youden diagram.

An ordered z-score chart is generated for the z-scores calculated for each test. On these charts each laboratory's z-score is shown, in order of magnitude, and is marked with its code number.

The advantages of these charts are that each laboratory is identified and the outliers are clearly indicated, however, unlike the Youden diagrams, they are not graphs of the actual results.

Youden two-sample diagrams are presented to highlight laboratory systematic differences. These charts are generated for pairs of results and are based on a plot of each laboratory's pair of results.

The advantages of these diagrams are that they are plots of the actual data - so the laboratories with results outside the ellipse can see *how* their results differ from the others - and results with an absolute z-score greater than 2.0 are highlighted.

It is important to note, however, that Youden diagrams are an illustration of the data only and are *not* used to assess the results (this is done by the z-scores).

These charts are to assist the Program Coordinator and Technical Advisers with the interpretation of the results and are very useful to participants - especially those participants with outliers because they can see how their results differ from those submitted by other laboratories.

APPENDIX C

EVALUATION PROCEDURES FOR CALIBRATION PROGRAMS

C.1 Introduction

This appendix outlines the procedures PTA uses to evaluate the results of its *calibration* programs and *measurement audit programs* (refer to Appendix B for procedures applicable to *testing* programs).

C.2 Calibration and Measurement Audit Programs

PTA uses the E_n number to evaluate each individual result from a laboratory participating in a calibration or measurement audit program. E_n stands for **E**rror **n**ormalised and for a result to be acceptable the E_n number should be between -1.0 and +1.0 i.e. $|E_n| \leq 1.0$. (The closer to zero the better.)

In *testing* interlaboratory comparisons a laboratory's z-score gives an indication of how close the laboratory's measurement is to the assigned value, however, in *calibration* interlaboratory comparisons and in measurement audits the E_n numbers indicate whether laboratories are within their particular measurement uncertainty of the reference value (assigned value).

The E_n numbers do not necessarily indicate which laboratory’s result is closest to the reference value. Consequently, laboratories reporting small uncertainties may have a similar E_n number to laboratories working to a much lower level of accuracy (i.e. larger uncertainties).

C.3 Graphical Displays for Calibration and Measurement Audit Programs

Graphs of reported results and their associated uncertainties are usually included in final reports for calibration and measurement audit programs.

It is important to note however that the graphs are an illustration of the data only and allow a broad comparison of all participants’ results/uncertainties. They do not represent an assessment of results (this is done by the E_n numbers).

C.4 Measurement Uncertainty (MU)

The measurement uncertainty reported by the laboratory is used in the calculation of the E_n number. The test items used in these programs usually have sufficient resolution, repeatability and stability to allow the laboratory to report an uncertainty equal to their claimed "*best measurement capability*".

End of Document