GUIDE TO PROFICIENCY TESTING AUSTRALIA

pt:a
proficiency testing australia

2016

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[2]. 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 - <u>testing</u> interlaboratory comparisons, which involve concurrent testing of samples by two or more laboratories and calculation of consensus values from all participants' results, and <u>calibration</u> 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 17025:2005 *General requirements for the competence of testing and calibration laboratories*

3. ISO/IEC 17011:2004 *Conformity assessment: General requirements for accreditation bodies accrediting conformity assessment bodies*

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

5. ISO 13528:2015 *Statistical methods for use in proficiency testing by interlaboratory comparisons*

6. APLAC PT001 (revised 2008) *Calibration interlaboratory comparisons*

7. APLAC PT002 (revised 2008) *Testing 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 this includes the determination of consensus values.
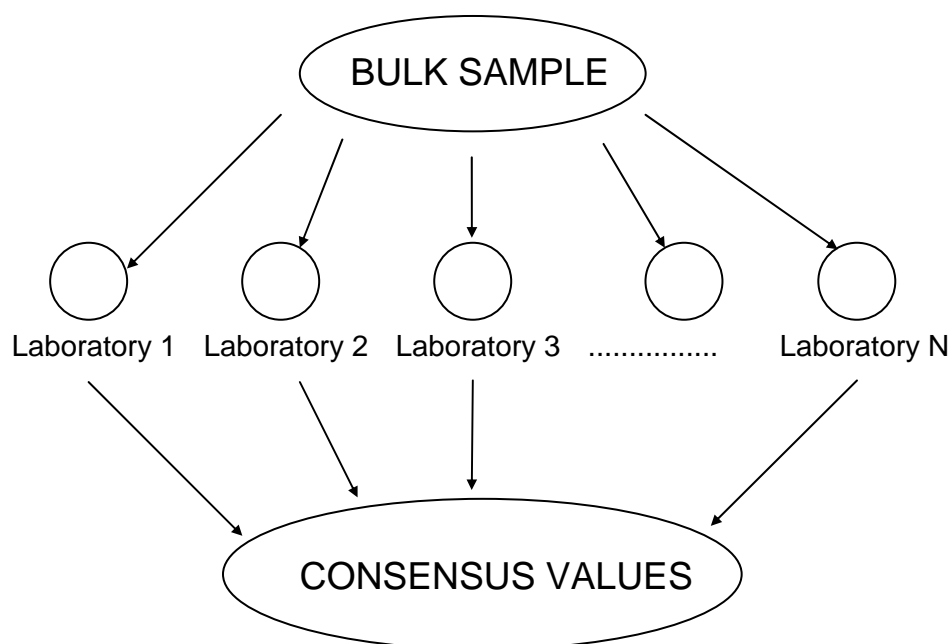

Figure 1: Typical Testing Interlaboratory Comparison

5.2 Working Group and Program Design

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

It is most important that at least one, but preferably two, technical experts are included in the planning of the program and in the evaluation of the results. Their input is needed in at least 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, answer questions from participants.

An appropriate statistical design is essential and therefore must be 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 each type of sample 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   Documentation

The main documents associated with the initial phase of a proficiency program are:

(a)   *Letter of Intent*

This is sent to prospective participants to advise that the program will be conducted and provides information on the type of samples and tests which will be included, the schedule and participation fees.

(b)   *Instructions to Participants*

These are carefully designed for each individual program and participants are always asked to adhere closely to them.

(c)   *Results Sheet*

For most programs a pro-forma results sheet is supplied to enable consistency in the statistical treatment of results.

Instructions and Results Sheets may be issued with, or prior to, the dispatch of samples.

5.5   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 this distribution. It is also ensured that certain restrictions on transport such as dangerous goods regulations or customs requirements are complied with.

5.6   Receipt of Results

Results from participating laboratories for PTA testing programs are required to be sent to either our Sydney office or Brisbane office. 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.

5.7   Analysis of Data and Reporting of Results

Results are usually analysed together (with necessary distinctions made for method variation) to give 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 six weeks of the due date for results.

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 (refer Appendix B) 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 to be 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 on the PTA website, and participants are notified of its availability via email.

5.8   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).

Further examples are one-off practical tests where material of known composition (e.g. certified reference material) is presented to one laboratory. This type of activity is also extensively used in the calibration area (refer Section 6.10, Measurement Audits). These activities do not, or by their nature cannot, use the usual consensus values as the basis for the evaluation of performance.

Some of PTA's 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 must also be 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.

Another difference between calibration and testing programs is that there is usually only one test item (also known as an artefact) which has to be distributed sequentially around the participating laboratories, making these programs substantially longer to run. Consequently, great care has to be taken to ensure the measurement stability of the test item.
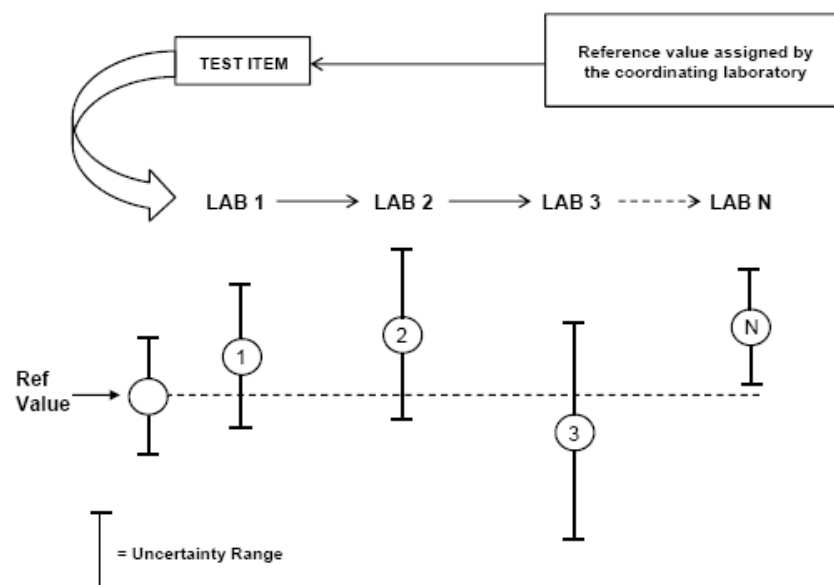

Figure 2: Typical Calibration Interlaboratory Comparison

In Figure 2, LAB 3 has a larger uncertainty range than LAB 1. This means that LAB 1 has the capability to calibrate higher accuracy instruments. This situation, where laboratories are working to different levels of accuracy, is valid provided that each laboratory works within their capabilities and that their nominated level of accuracy (measurement uncertainty) is suitable for the instrument being calibrated.

6.2 Program Design

Once a program has been selected, a small working group is formed. This group usually comprises one or more Technical Advisers and a PTA Scientific Officer 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. For example, it would be inappropriate to send a 3½ digit multimeter to a laboratory that had a nominated measurement uncertainty of 5 parts per million (0.0005%) because the resolution, repeatability and stability of such a test item would limit the measurement uncertainty the laboratory could report to no better than 0.05%. What is necessary is 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.

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 Documentation

A *Letter of Intent* is sent to all potential participants to advise that the program will be conducted and to provide as much information as possible.

Instructions to Participants are carefully designed for each individual program and it is essential to the success of the program that the participating laboratories adhere closely to them. For most programs a pro-forma *Results Sheet* is used, to ensure that laboratories supply all the necessary information in a readily accessible format.

6.5 Test Item Stability

The test item is distributed sequentially around the participating laboratories. To ensure its stability, it is usually calibrated at least at the start and at the end of the circulation. For test items whose values may drift during the course of the program (e.g. resistors, electronic devices, etc.) more frequent calibrations and checks are necessary.

6.6 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.

6.7   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.

As stated previously, it is important to select test items with high resolution, good repeatability and good stability. This is to ensure that these factors do not contribute significantly to the reference value uncertainty. Likewise, the Reference Laboratory must have the capability to assign measurement uncertainties that are better than the participating laboratories. Otherwise it will be more difficult to evaluate each laboratory's performance.

Where a test item has exhibited drift, the reference values will usually be derived from the mean of the Reference Laboratory calibrations carried out before and after the measurements made by the participating laboratories. Where a step change is suspected, then the reference values will be derived from the most appropriate Reference Laboratory calibration.

6.8   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.

6.9   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 on the PTA website (www.pta.asn.au) at the conclusion of the program. This typically contains more information than is provided in the summary sheet - including all participant's results and uncertainties, final E_n numbers, technical commentary and graphical displays.

6.10 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 (usually supplied by NMI).

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

APPENDIX A

GLOSSARY OF TERMS

<u>GLOSSARY OF TERMS</u>

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[1].

assigned value	value attributed to a particular property of a proficiency test item
consensus value	an assigned value obtained from the results submitted by participants (e.g. for most testing programs the median[†] is used as the assigned value)
E_n number	stands for error normalised and is the internationally accepted quantitative measure of laboratory performance for calibration programs (see formula in Appendix C)
false negative	failing to report the presence of a parameter (e.g. analyte, organism) which is present in the sample
false positive	erroneously reporting the presence of a parameter (e.g. analyte, organism) which is absent from the sample
interlaboratory comparison	organisation, 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	observation in a set of data that appears to be inconsistent with the remainder of that set, e.g. absolute z-score greater than or equal to three (i.e. 3.0) for testing programs
reference value	an assigned value which is provided by a Reference Laboratory
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 - e.g. (result – median[†]) ÷ normalised IQR[†]

NOTE: [†] *the median, normalised interquartile range (IQR) and other summary statistics are defined in Appendix B.*

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 3).
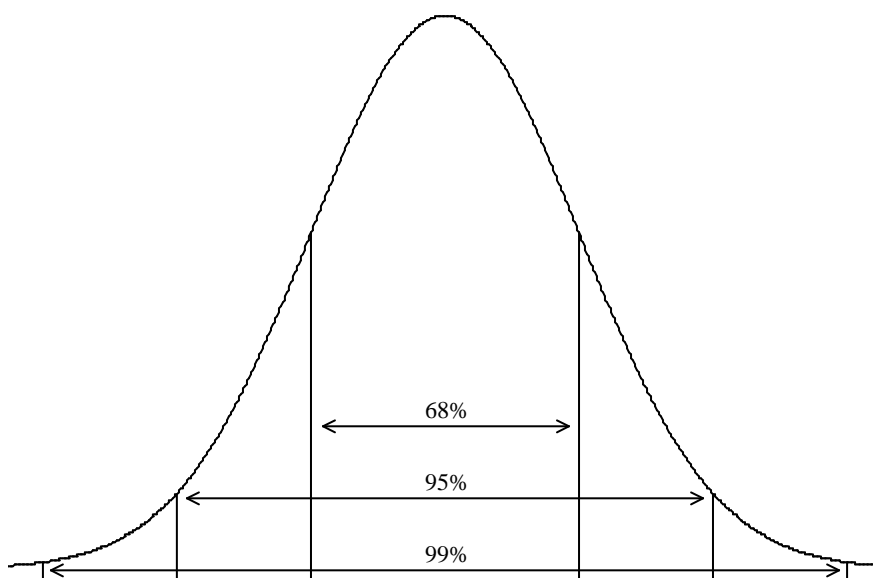
Figure 3: The Normal Distribution

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 will be approximately normal the working group (in particular the Technical Adviser) must think carefully about 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 of the 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 of the 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. One problem which may arise is that there are two distinct groups of results 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 eight summary statistics - number of results, median, uncertainty of the median, normalised interquartile range (IQR), robust coefficient of variation (CV), minimum, maximum and range. All of these are described in detail below.

The most important statistics used are the median and the normalised IQR - these are measures of the centre and spread of the data (respectively), similar to the mean and standard deviation. The median and normalised IQR are used because they are robust statistics, which means that they are not influenced by the presence of outliers in the data.

The no. of results is simply the total number of results received for a particular test/sample, and is denoted by N. Most of the other statistics are calculated from the sorted results, i.e. from lowest to highest, and in this appendix X[i] will be used to denote the ith sorted data value (e.g. X[1] is the lowest value and X[N] is the highest).

The median is the middle value of the group, i.e. half of the results are higher than it and half are lower. If N is an odd number the median is the single central value, i.e. X[(N+1)/2]. If N is even, the median is the average of the two central values, i.e. (X[N/2] + X[(N/2)+1])/2. For example if N is 9 the median is the 5th sorted value and if N is 10 the median is the average of the 5th and 6th values.

The normalised IQR is a measure of the variability of the results. It is equal to the interquartile range (IQR) multiplied by a correction factor[†], which makes it comparable to a standard deviation. The interquartile range is the difference between the lower and upper quartiles. The lower quartile (Q1) is the value below which, as near as possible, a quarter of the results lie. Similarly the upper quartile (Q3) is the value above which a quarter of the results lie. In most cases Q1 and Q3 are obtained by interpolating between the data values. The IQR = Q3 − Q1 and the normalised IQR = IQR × correction factor.

Since the median is a consensus value, it has an uncertainty originating from the testing conditions of the laboratories that participated in the program and other factors. The (standard) uncertainty of the median is calculated as:

$$\text{uncertainty(median)} \approx \sqrt{\frac{\pi}{2}} \times \frac{\text{normalised IQR}}{\sqrt{N}}$$

where N = no. of results.

The robust CV is a coefficient of variation (which allows for the variability in different samples/tests to be compared) and is equal to the normalised IQR divided by the median, expressed as a percentage - i.e. robust CV = 100 × normalised IQR ÷ median.

The minimum is the lowest value (i.e. X[1]), the maximum is the highest value (X[N]) and the range is the difference between them (X[N]–X[1]).

On page 17 is an example of the summary statistics as they appear in a final report.

NOTE: [†] *The interquartile range of normally distributed data is not equivalent to the familiar ±1 SD interval. To convert an IQR into a ±1 SD range, it must be scaled by a correction factor. The correction factor is calculated by using expected normal scores of order statistics and depends on the number of results reported for the test/sample.*

Example: Data Set and Summary Statistics

Waters (Chemical) Results for PTA Sample 1 - Total Solids, Total Suspended Solids and Total Dissolved Solids

Lab Code	PTA Sample 1						Total Solids Robust Z-Scores	Total Suspended Solids Robust Z-Scores	Total Dissolved Solids Robust Z-Scores
	Total Solids		Total Suspended Solids		Total Dissolved Solids				
	Result ± MU mg/L		Result ± MU mg/L		Result ± MU mg/L				
1	584	25	200	6	389	25	-0.91	-0.51	-1.64
2	600	60	204	#	405	40	-0.29	-0.26	-0.10
3	572	15	195	20	406	20	-1.37	-0.84	0.00
4	624	#	216	#	431	#	0.64	0.51	2.41
5	575	#	192	10	444	#	-1.25	-1.03	3.66 §
6	631	113	209	#	410	#	0.91	0.06	0.39
7	640	64	176	#	351	8.6	1.25	-2.06	-5.30 §
8	600	1	180	#	360	36	-0.29	-1.80	-4.43 §
9	581	58.1	185	7.6	410	41	-1.02	-1.48	0.39
10	592	#	190	34	432	1	-0.60	-1.16	2.51
11	567.5	#	230	23	395	39.5	-1.54	1.41	-1.06
12	621	13	222	1	410	#	0.52	0.90	0.39
13	602	#	181	18.1	370	#	-0.21	-1.73	-3.47 §
14	625	63	182	#	426	10	0.67	-1.67	1.93
15	620	8.37	195	#	368	#	0.48	-0.84	-3.66 §
16	611	16.74	223	1.7	413	41	0.13	0.96	0.67
17	586	#	226	#	407	7.53	-0.83	1.16	0.10
18	627	30	201	20	402	4.21	0.75	-0.45	-0.39
19	619	40	213	10.58	396	#	0.44	0.32	-0.96
20	700	#	214	5.79	408	20	3.57 §	0.39	0.19
21	600	6.28	178	#	398	60	-0.29	-1.93	-0.77
22	624	64.90	207	15	409	6.13	0.64	-0.06	0.29
23	588	#	209	15	406	28.42	-0.75	0.06	0.00
24	619	31.7	211	21	405	#	0.44	0.19	-0.10
25	634	15	203	3.02	410	20.5	1.02	-0.32	0.39
26	624	10	218	27.47	390	59	0.64	0.64	-1.54
27	604	72	226	32.3	396	47.5	-0.13	1.16	-0.96
28	578	58	182	#	411	15	-1.14	-1.67	0.48
29	601	60	213	6	404	8.8%	-0.25	0.32	-0.19
30	<500	40	216	15.1	419	10	0.98	0.51	1.25

NOTES: § denotes an outlier, i.e. |z-score|≥3.0.
# indicates that no results were submitted.
"N/A" indicates not applicable.

**TOTAL SOLIDS, TOTAL SUSPENDED SOLIDS AND TOTAL DISSOLVED SOLIDS -
SUMMARY STATISTICS (mg/L)**

Statistic	Total Solids	Total Suspended Solids	Total Dissolved Solids
No. of Results	30	30	30
Median	607.5	205.5	406.0
Normalised IQR	25.9	18.5	10.4
Uncertainty (Median)	5.9	4.2	2.4
Robust CV	4.3%	9.0%	2.6%
Minimum	567.5	176	351
Maximum	700	230	444
Range	132.5	54	93

B.5   Robust Z-scores and Outliers

To statistically evaluate the participants' results, PTA uses z-scores based on robust summary statistics (the median and normalised IQR).

If a sample in a testing program is labelled A, then the robust z-score (denoted by Z) for a laboratory's sample A result would be:

$$Z = \frac{A - \text{median}(A)}{\text{normIQR}(A)}$$

where the median and normalised IQR of all the sample A results are denoted by median(A) and normIQR(A), respectively.

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 <u>outlier</u> 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 the example on page 16, laboratory 5 has a positive outlier for Total Dissolved Solids and laboratory 20 has a positive outlier for Total Solids. Laboratories 7, 8, 13 and 15 have negative outliers for Total Dissolved Solids.

In some circumstances it may not be possible to calculate a robust z-score using the formula above. This occurs when the normalised IQR is equal to zero (which could occur if more than 50% of the results submitted by participants were identical and equal to the median). In other circumstances it may be possible to calculate a robust z-score using the formula above, but the spread of results (as measured by the normalised IQR) might be so small that even a slight deviation from the median will result in an outlier. In yet other circumstances the spread of results (as measured by the normalised IQR) might be so large that it is extremely unlikely that any result would ever be classified as an outlier.

If the normalised IQR is equal to zero, or if the spread of results is too large or too small, in the opinion of the Technical Adviser, then a target coefficient of variation (CV) is used to calculate z-scores. These z-scores are calculated by:

$$Z = \frac{A - \text{median}(A)}{\text{target CV} \times \text{median}(A)}$$

where the target CV is expressed as a decimal.

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.

Suppose the pair of results are from two samples labelled A and B. The standardised sum (denoted by S) and standardised difference (D) for the pair of results are:

$$S = (A + B) / \sqrt{2} \quad \text{and} \quad D = \begin{cases} (B - A) / \sqrt{2}, & \text{if median}(A) < \text{median}(B) \\ (A - B) / \sqrt{2}, & \text{otherwise.} \end{cases}$$

Each laboratory's standardised sum and difference are calculated, followed by the median and normalised IQR of all the S's and all the D's - i.e. median(S), normIQR(D), etc.

The between-laboratories z-score (denoted by ZB) is then calculated as the robust z-score for S and the within-laboratory z-score (ZW) is the robust z-score for D, i.e.

$$ZB = \frac{S - \text{median}(S)}{\text{normIQR}(S)} \quad \text{and} \quad ZW = \frac{D - \text{median}(D)}{\text{normIQR}(D)}$$

B.6 Graphical Displays

In addition to tables of the results and z-scores, and summary statistics, 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 - both of which are described in detail below.

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.

Ordered Z-score Chart

An ordered z-score chart is generated for the z-scores calculated for each test. An example is included below. On these charts each laboratory's z-score is shown, in order of magnitude, and is marked with its code number. From this each laboratory can readily compare its performance relative to the other laboratories.

These charts contain solid lines at +3.0 and -3.0, so the outliers are clearly identifiable as the laboratories whose "bar" extends beyond these cut-off lines. The y-axis is usually limited, so in some cases very large or small (negative) z-scores appear as extending beyond the limit of the chart - for example, laboratory 7 for the Total Dissolved Solids z-score bar-chart on page 20.
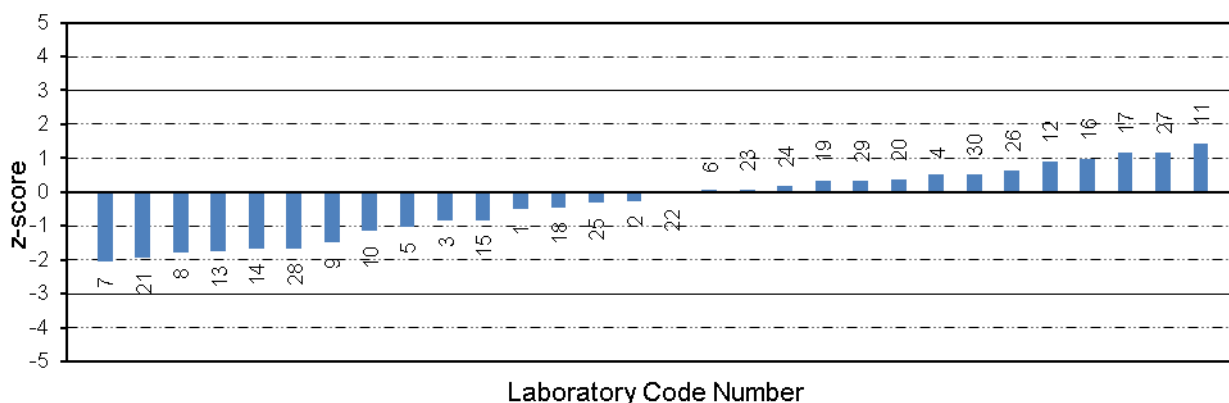
The advantages of these charts are that each laboratory is identified and the outliers are clearly indicated, however, unlike the Youden diagrams, they are not graphs of the actual results.
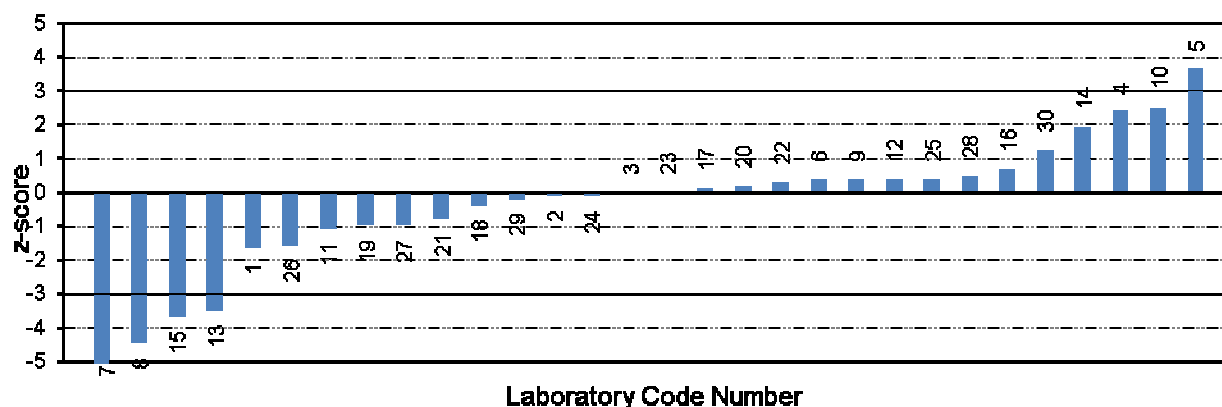
Examples: Ordered Z-Score Charts

Total Solids - Sample PTA 1 - Robust Z-Scores
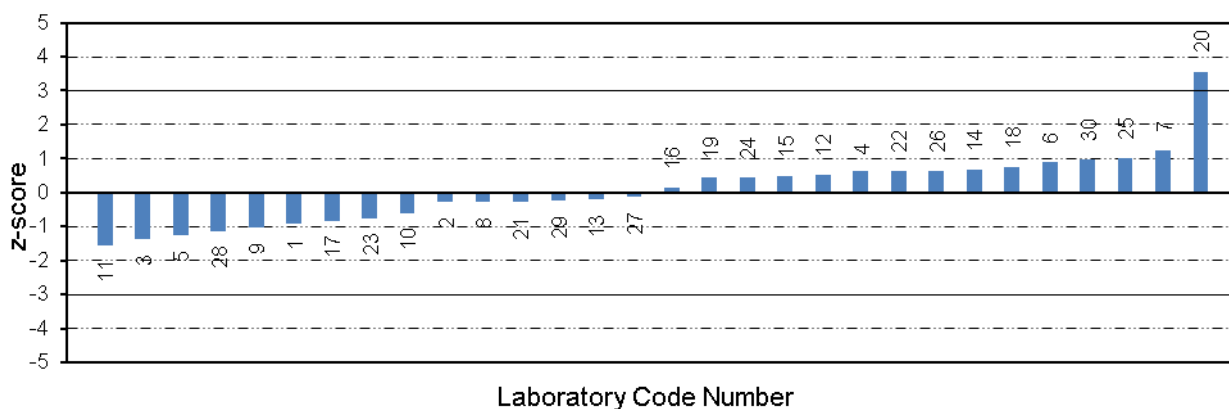
Total Suspended Solids - Sample PTA 1 - Robust Z-Scores

Total Dissolved Solids - Sample PTA 1 - Robust Z-Scores

Youden Diagrams

These charts are generated for pairs of results. Youden diagrams are produced for biological program reports where results have been log transformed, for duplicate samples, and for duplicate results requested from the same sample. Youden two-sample diagrams are presented to highlight laboratory systematic differences. They are based on a plot of each laboratory's pair of results, represented by a black spot •.

These diagrams also feature an approximate 95% confidence ellipse for the bivariate analysis of the results, and dashed lines which mark the median value for each of the samples. The ellipse is estimated by re-scaling an approximate 95% confidence region (which is a circle) in the bivariate z-scores space back to the original data space.

All points which lie outside the ellipse are labelled with the corresponding laboratory's code number. Note, however, that these points may not correspond with those identified as outliers. This is because the outlier criterion ($|Z| \geq 3.0$) has a confidence level of approximately 99%, whereas the ellipse is an approximate 95% confidence region.

This means that, if there are no outliers in the data, it can be expected that about 5% (i.e. one in twenty) of the results will lie outside the ellipse, however, as proficiency testing data usually contains some outliers, more than 5% of points will be outside the ellipse in most cases. The points outside the ellipse on the Youden diagram will roughly correspond to those with absolute z-scores greater than 2.0. Laboratories with results outside the ellipse which have not been identified as outliers (those which have $2.0 < |Z| < 3.0$) are encouraged to review their results.

An example of a Youden diagram is included below. All of the laboratories with outliers, i.e. $|Z| \geq 3.0$, and those with $2.0 < |Z| < 3.0$ lie outside the ellipse.

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.

As a guide to the interpretation of the Youden diagrams:

(i) laboratories with significant systematic error components (i.e. between-laboratories variation) will be outside the ellipse in either the upper right hand quadrant (as formed by the median lines) or the lower left hand quadrant, i.e. inordinately high or low results for *both* samples;

and

(ii) laboratories with random error components (i.e. within-laboratory variation) significantly greater than other participants will be outside the ellipse and (usually) in either the upper

left or lower right quadrants, i.e. an inordinately high result for one sample and low for the other.

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).

Example: Youden diagram

Aerobic Plate Count, All Methods Pooled - log(cfu/g)
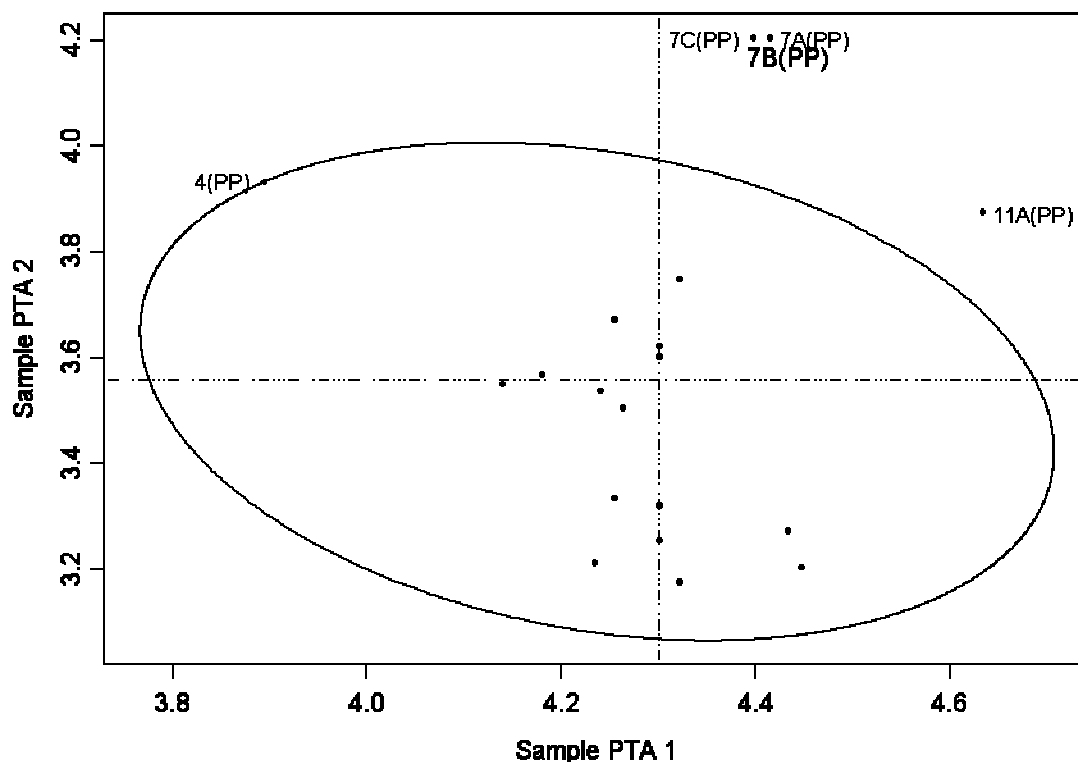
B.7   Laboratory Summary Sheets

In addition to the final report, which contains complete details of the statistical analysis, an individual summary sheet is prepared for each participant. This laboratory summary sheet contains all of the participant's results, alongside the statistics for that test/sample and the associated z-scores. Comments about the program in general and specific to the laboratory (if necessary) are also included.

An example summary sheet appears on page 23. At the top of the page is the title of the program and the identity of the laboratory. The main part of this summary sheet consists of: the test and sample identity; the laboratory's result including its MU (where required); the number of results; median and normalised IQR for each test/sample; and the z-scores (or two z-scores for a sample pair) for each test.

Any outliers are again marked with a **§** next to the z-score. At the bottom of the page is a section for notes and comments. In this case there are no special laboratory-specific remarks. From this summary sheet we can see quickly and easily that:

(1)   this laboratory submitted results for all of the tests;

(2)   the laboratory has reported one outlier; and

(3)   the laboratory has reported one questionable result.

Seeing all of a laboratory's z-scores together can be very useful, even if no outliers were reported. For example, where a pair of samples is tested, if all of the between-laboratories z-scores are negative (or positive) this may be indicative of a laboratory bias - i.e. all of its results are lower (or higher) than the consensus values.

Example: Summary Sheet

Proficiency Testing Australia

LABORATORY SUMMARY SHEET

Proficiency Testing - Waters (Chemical) Round [###]
- Total Solids, Total Suspended Solids, Total Dissolved Solids -
Report No. [###]

Date of summary sheet issue: [Date]

Lab Name:	*[name of Laboratory/company, including Site name]*	**Laboratory Code:**	[##]
Location:	*[state/country]*		

Analyte	Sample	Laboratory result ± MU (mg/L)[1]	Median[2]	Norm. IQR[3]	Robust CV[4]	No. of results	Robust z-score[5]	
Total Solids (TS)	PTA 1	640 ± 64	607.5	25.9	4.3%	30	1.25	
Total Suspended Solids (TSS)	PTA 1	176 ± ▪	205.5	18.5	9.0%	30	-2.06	**?**
Total Dissolved Solids (TDS)	PTA 1	351 ± 9	406.0	10.4	2.6%	30	-5.30	**§**

No. of outlier results is: 1

[1]   A "▪" indicates that no result was returned for this sample/test.

[2]   The *median* is the middle result. It is a measure of the centre of the data set.

[3]   The *normalised IQR* is a measure of the spread of the results. It is calculated by multiplying the *interquartile range* (*IQR*) by a factor which converts the *IQR* to an estimate of the *standard deviation*. The *IQR* is the difference between the *upper* and *lower quartiles* (i.e. the values above and below which a quarter of the results lie, respectively).

[4]   The *robust coefficient of variation* (*robust CV*) is calculated by dividing the *normalised IQR* by the *median* and expressed as a percentage. The *robust CV* allows for the variability in different samples/tests to be compared.

[5]   Each *z-score* marked with a "**§**" is an *outlier* (i.e. |*z-score*| ≥ 3.0). Laboratories are also encouraged to review results which have an absolute *z-score* value between two and three (i.e. 2.0 < |*z-score*| < 3.0), these have been marked with a "?".

[6]   For the purposes of consistency in reporting, summary sheet results and MU values have been rounded to zero decimal places for all analytes.

This summary sheet should be read in conjunction with the final report found at www.pta.asn.au. The above results are from one proficiency program only and may not be fully representative of a laboratory's overall performance. Therefore, this summary sheet should not be used solely to evaluate laboratory competence.

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). The procedures used by PTA are consistent with those used for international calibration programs run by the European Cooperation for Accreditation (EA) and Asia Pacific Laboratory Accreditation Cooperation (APLAC).

C.2 Calibration Program

As stated in Section 6.6, PTA uses the E_n number to evaluate each individual result from a laboratory. E_n stands for **E**rror **n**ormalised and is defined as:-

$$E_n = \frac{LAB - REF}{\sqrt{U^2{}_{LAB} + U^2{}_{REF}}}$$

> where: LAB is the participating laboratory's result
> REF is the Reference Laboratory's result
> U_{LAB} is the participating laboratory's reported uncertainty
> U_{REF} is the Reference Laboratory's reported uncertainty

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 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, calibration 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).

In a series of similar measurements a normal distribution of E_n numbers would be expected. So when considering the significance of any results with $|E_n|$ marginally greater than 1.0, all the results from that laboratory are evaluated to see if there is a systematic bias e.g. consistently positive or consistently negative values of E_n.
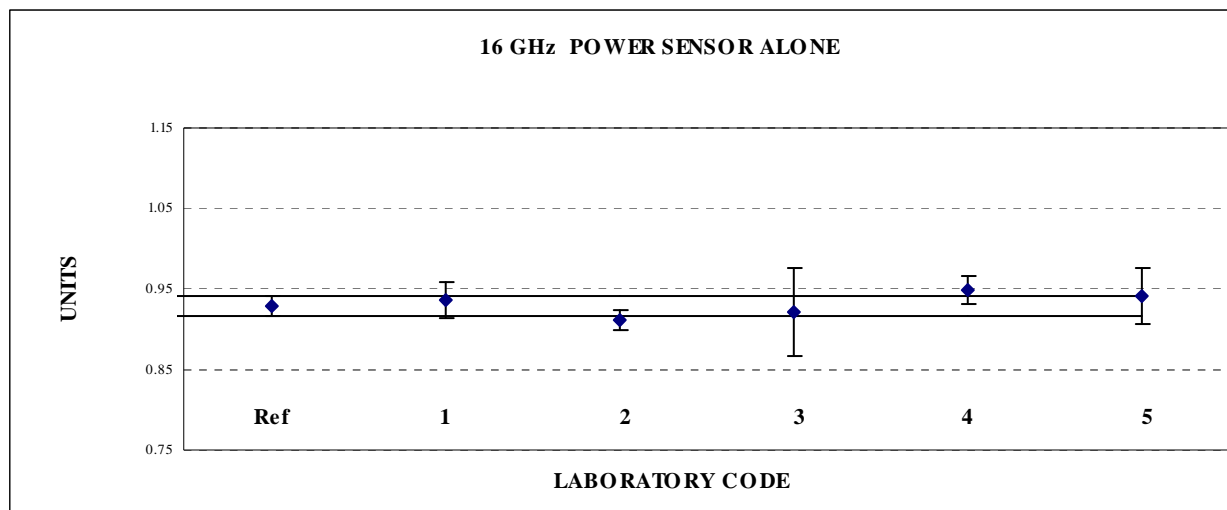
A sample of results from a radio frequency power interlaboratory comparison, their corresponding reported uncertainties and E_n numbers are tabulated below. The result for laboratory 2 is considered unsatisfactory.

16 GHz Power Sensor Alone

Lab Code	Results	U_{95}	E_n
REF	0.929	0.011	
1	0.936	0.022	0.28
2	0.911	0.012	**-1.09**
3	0.921	0.054	-0.14
4	0.949	0.018	0.94
5	0.942	0.035	0.35

C.3 Graphical Displays for Calibration Program

Graphs of reported results and their associated uncertainties are included in final reports for *calibration* programs. The example graph below shows a plot of the results tabulated in Section C.2. Each laboratory's result is represented by a ◆ mark. The bars protruding above and below the ◆ mark represent that laboratory's reported measurement uncertainty, that is, the region in which the laboratory has statistically calculated (with a 95% confidence level) that the "true value" may lie, or in other words, their estimate of how accurately they can measure.

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 Audit Programs

A sample of results from a pressure transducer *measurement audit*, the laboratory's corresponding reported uncertainties and E_n numbers are tabulated below. The results for decreasing applied pressures at 9.9999 MPa, 7.5000 MPa and 5.0000 MPa are considered unsatisfactory.

10 MPa Pressure Transducer

APPLIED PRESSURE	REF VALUE MPa	REF U$_{95}$ MPa	LAB MEAN MPa	LAB U$_{95}$ MPa	E$_n$ NO.
5.0000	4.8983	0.0014	4.8982	0.002	-0.03
7.5000	7.3478	0.0014	7.3466	0.002	-0.46
9.9999	9.7973	0.0019	9.7970	0.004	-0.08
9.9999	9.8133	0.0025	9.7972	0.004	**-3.72**
7.5000	7.3605	0.0031	7.3462	0.002	**-3.88**
5.0000	4.9074	0.0025	4.8971	0.002	**-3.51**

Graphs of reported results and their associated uncertainties are provided for *measurement audit* programs when necessary.

C.5   Measurement Uncertainty (MU)

The measurement uncertainty reported by the laboratory is used in 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