



DPLP 10 Rev. 0

KEBIJAKAN KAN MENGENAI TATA CARA UJI PROFISIENSI

Komite Akreditasi Nasional
National Accreditation Body of Indonesia
Gedung Manggala Wanabakti, Blok IV, Lt. 4
Jl. Jend. Gatot Subroto, Senayan, Jakarta 10270 – Indonesia
Tel. : 62 21 5747043, 5747044
Fax. : 62 21 57902948, 5747045
Email : laboratorium@bsn.or.id
Website : <http://www.bsn.or.id>

Dokumen ini tidak dikendalikan jika di-download/Uncontrolled when downloaded

CONTENTS

| | <i>Page</i> |
|---|-------------|
| 1. Scope | 2 |
| 2. Introduction | 2 |
| 3. Testing Interlaboratory Comparisons | 2 |
| 3.1 Introduction | 3 |
| 3.2 Working Group and Program Design | 3 |
| 3.3 Sample Supply and Preparation | 3 |
| 3.4 Documentation | 3 |
| 3.5 Packaging and Dispatch of Samples | 4 |
| 3.6 Receipt of Results | 4 |
| 3.7 Analysis of Data and Reporting of Results | 4 |
| 3.8 Evaluation of Performance and Follow-Up | 5 |
| 4. Calibration Interlaboratory Comparisons | 5 |
| 4.1 Introduction | 5 |
| 4.2 Program Design | 6 |
| 4.3 Artefact Selection | 6 |
| 4.4 Documentation | 6 |
| 4.5 Artefact Stability | 7 |
| 4.6 Evaluation of Performance | 7 |
| 4.7 Reference Values | 7 |
| 4.8 Uncertainty of Measurement | 8 |
| 4.9 Reporting | 8 |
| 4.10 Corrective Action | 9 |
| 4.11 Measurement Audits | 9 |
| 5. References | 9 |
| Appendix A Evaluation Procedures for Testing Programs | 10 |
| Appendix B Evaluation Procedures for Calibration Programs | 16 |

1. Scope

The purpose of this document is to provide participants in KAN's proficiency testing programs, in particular KAN accredited and applicant laboratories, with an overview of how the various types of proficiency testing programs are conducted by KAN 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 KAN's internal procedures which are in compliance with the requirements of ISO/IEC Guide 43.1.

The main body of this document contains general information about KAN's proficiency testing 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

KAN's proficiency testing is 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.

Proficiency testing is carried out by KAN's Proficiency Testing Team (PTT) consisting of statisticians and scientists with experience in proficiency testing with technical input for each program provided by honorary technical advisers. KAN PTT is responsible to preparation of proficiency testing protocol, to carry out and to evaluate the result.

3. Testing Interlaboratory Comparisons

3.1 Introduction

Most of interlaboratory comparisons conducted by KAN subdivided samples (taken from a bulk sample) are distributed to participating laboratories which test them concurrently. They then return results to KAN for analysis and this includes the determination of consensus values.

3.2 Working Group and Program Design

KAN proficiency testing is carried out annually. The program for each year is selected by KAN PTT at the beginning of the respective year.

The preparation of program includes:

- 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 significant figures/decimal places to which results should be reported and correct units for reporting;
- technical commentary in the final report and in some cases evaluation of replies submitted by laboratories which were requested to investigate extreme results.

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

3.3 Sample Supply and Preparation

Sample for KAN proficiency testing is prepared by provider laboratories. Criteria of sample/ artifact provider for being appointed is as the following:

- be KAN-accredited laboratory for testing laboratory and National Measurement Institute for calibration laboratory;
- having sufficient competency in the preparation of test sample and/ or artifact.

3.4 Documentation

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

(a) *Letter of Intent*

This is sent to prospective participants, including all accredited and applicant laboratories 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.

3.5 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. Certain restrictions on transport such as dangerous goods regulations or customs requirements are complied with.

3.6 Receipt of Results

Results from participating laboratories for KAN's proficiency testing programs are required to be sent to Secretariat of KAN. 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 may not be included in the data analysis, laboratories are strongly encouraged to submit all results on time.

3.7 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 an interim report can be issued to participants within two to four weeks of the due date for results.

This letter includes preliminary feedback including the consensus values for each test/sample and also features of the program's design. Laboratories reporting one or more result(s) that are significantly different from the consensus values are encouraged to commence investigative/corrective action prior to the release of the final report.

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 A.

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 A)
- (e) a table summarising the extreme† results
- (f) Technical comments (on possible causes of extreme results, method effects overall performance etc.)
- (g) sample preparation and homogeneity testing information
- (h) a copy of the instructions to participants and results sheet

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

3.8 Evaluation of Performance and Follow-up

Laboratories reporting extreme results for accredited tests are requested to investigate and report back to KAN. This request is made in a covering letter

which accompanies the final report and laboratory summary sheet. Each laboratory's response is then reviewed during next on site assessment.

4. Calibration Interlaboratory Comparisons

4.1 Introduction

Each calibration laboratory accredited by KAN has its accredited capability uniquely expressed both in terms of its ranges of measurements and the best measurement capability 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 are accredited (their *least uncertainties of measurement*).

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 KAN interlaboratory comparisons the Reference Laboratory is usually the KIM-LIPI, which maintains Indonesia's primary standards of measurement.

Another difference between calibration and testing programs is that there is usually only one test item (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 artefact.

4.2 Program Design

Once a program has been selected, a small working group is formed. This group usually comprises one or more technical advisers from KIM-LIPI as NMI, consist of measurement report for the related field KAN proficiency testing officer who will act as the program coordinator. The group decides on the measurements to be conducted, how often the artefact will need to be recalibrated and the range of values to be measured. They also formulate instructions and results sheets. KAN's program are designed so that it will normally take no more than one year for each program.

4.3 Artefact Selection

Because there can often be a substantial difference in the accredited uncertainties of measurements of the participating laboratories the artefact must be carefully chosen. For example, it would be inappropriate to send a 3½ digit multimeter to a laboratory that had an accredited uncertainty of measurement of 5 parts per

million (0.0005%) because the resolution, repeatability and stability of such an artefact would limit the uncertainty of measurement the laboratory could report to no better than 0.05%. What is necessary is an artefact 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.

4.4 Documentation

A *Letter of Intent* is sent to all accredited and applicant laboratories 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.

4.5 Artefact Stability

The artefact 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 artefacts whose values may drift during the course of the program (e.g. resistors, electronic devices, etc.) more frequent calibrations and checks are necessary.

4.6 Evaluation of Performance

As stated in Section 7.1, 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 intercomparison. Instead, their results are compared only to the Reference Laboratory's results and their ability to achieve the accuracy for which they are accredited is evaluated by calculating the En ratio. For further details please refer to Appendix B.

4.7 Reference Values

KIM-LIPI provides most of the reference values for KAN interlaboratory comparisons. The majority of the participating laboratories' reference equipment is also calibrated by KIM-LIPI.

As stated previously, it is important to select an artefact 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 uncertainties of measurement that are better than the participating laboratories. Otherwise it will be more difficult to evaluate each laboratory's performance.

Where an artefact 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.

Where an artefact is calibrated a number of times during an intercomparison, participants' results can be graphed using the *difference* between each laboratory's result and the appropriate reference value (LAB-REF), rather than the actual laboratory results and corresponding reference values (refer Appendix C). This also helps maintain the confidentiality of the reference values for later use of the artefact.

4.8 Uncertainty of Measurement

To be able to adequately intercompare 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 Guide4, which has been adopted by laboratory it specific field.

Laboratories should not report uncertainties smaller than their accredited uncertainty of measurement.

4.9 Reporting

Whenever practicable an *Interim Report* is sent to laboratories to give them early feedback on their performance.

The interim report states the En values for each measurement based on the preliminary reference values and usually does not contain any technical commentary.

A *Final Report* is sent to the Authorised Representative of each laboratory at the conclusion of the program. This typically contains more information than is provided in the interim report - including all participant's results and uncertainties, final En numbers, technical commentary and graphical displays.

4.10 Corrective Action

Where a laboratory reports results which are deemed unsatisfactory (refer Appendix C) they will be notified, asked to check their results and provide an explanation. If it is a measurement problem then every effort will be made to arrange a re-test of the artefact. The reference values will not be released to the laboratory until after the retest results have been received, however, they may be given an indication of which results require attention and whether they are consistently higher or lower than the reference values.

If the problem is not satisfactorily resolved, further action as outlined in Section 5.2 will then be considered in conjunction with the Technical Adviser and the Field Manager.

7.11 Measurement Audits

The term *measurement audit* is used by KAN to describe a practical test whereby a well characterised and calibrated test item (artefact) is sent to a laboratory and the results are compared with a reference value (usually supplied by KIM-LIPI). Where suitable artefacts exist, these audits are mandatory for calibration laboratories applying for accreditation or significantly extending their scope of accreditation. They may also be employed for the evaluation of proposed KAN signatories or where laboratories due for reassessment have not recently participated in a proficiency testing activity in a particular area.

Normally, the artefact(s) will be sent to a laboratory prior to an assessment so that the resulting information is available to the assessment team. In some cases the measurement audit (or part of it) is witnessed by the assessment team, for instance, when a proposed signatory is carrying out the measurement.

Procedures are the same as for a normal interlaboratory comparison except that usually only a simple report is generated. In some cases the laboratories may only be told that they are within their uncertainty of the reference value.

If problems arise the laboratory is given the chance to retest, but subsequent poor performance would preclude granting of accreditation for specific calibration area.

5. References

1. ISO/IEC Guide 43: 1997 *Proficiency testing by interlaboratory comparisons - Part 1: Development and operation of proficiency testing schemes Part 2: Selection and use of proficiency testing schemes by laboratory accreditation bodies*
2. ISO/IEC 17025 : 1999 *General requirements for the competence of testing and calibration laboratories*
3. ISO/IEC Guide 58: 1993 *Calibration and testing laboratory accreditation systems - General requirements for operation and recognition*
4. ISO *Guide to the expression of uncertainty in measurement*: Corrected and reprinted, 1995 (BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML)
5. EA 2/03 November 2001 *EA Interlaboratory Comparisons*
6. APLAC PT001 (revised 1999) *Calibration interlaboratory comparisons*
7. APLAC PT002 (revised 1998) *Testing interlaboratory comparisons*
8. ILAC-G13:2000 *Guidelines for the Requirements for the Competence of Proficiency Testing Schemes*

APPENDIX A

EVALUATION PROCEDURES FOR TESTING PROGRAMS

A.1 Introduction

A.2 Statistical Design

A.3 Data Preparation

A.4 Summary Statistics

A.5 Robust Z-scores & Outliers

A.6 Graphical Displays

A.1 Introduction

This appendix outlines KAN procedures 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 B). 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 procedures can be used to analyse the results. However in some cases a program is run where the common statistical analyses cannot be applied - in these cases other, more appropriate, statistical procedures may be used.

Sections A.4, A.5 and A.6 of this appendix outline the actual statistical analysis used (including some examples) - i.e. the statistics, tables and charts which appear in the final report for the program. The next section (A.2) examines some background theory, which is considered during the planning of a program, and Section A.3 describes the collection, entry and checking of results which are carried out prior to commencing the statistical analysis.

A.2 Statistical Design

Given that any differences between the samples have been minimised, variability in the results for a program has two main sources - variation between laboratories (which may include variation between methods) and variation within a laboratory. It is desirable to evaluate and provide feed-back on both of these types of variation.

In order to assess both between-laboratories and within-laboratory variation, participants must perform the same testing more than once (e.g. twice). Therefore, programs are designed such that, whenever possible, pairs of related results are obtained. This can be achieved by using pairs of related samples or, if this is not possible, by requesting two results on one sample.

If paired samples are used they may be identical 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 statistical analysis of the results is the same for both types of pairs (uniform or split), but the interpretation is slightly different (see Section A.5). For some programs it is not possible to achieve pairs of results - i.e. a single result on a single sample is obtained. In this case the statistical analysis is simpler, but it is not possible to differentiate between the two types of variation.

The other main statistical consideration 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.

A.3 Data Preparation

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

As the results are submitted to KAN 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 log₁₀ 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.

A.4 Summary Statistics

Once the data preparation is complete, summary statistics are calculated to describe the data. KAN employes seven summary statistics - number of results, 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 i th 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 factor† (0.7413), 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 = $IQR \times 0.7413$.

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 = $100 \times \text{normalised IQR} \div \text{median}$.

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 ($X[N]-X[1]$).

Once the summary statistics have been calculated for each of the samples and tests in a program, the medians and normalised IQRs are tabulated and sent to each laboratory which has returned results as "early information". Following the issue of this information no further changes or additions (e.g. of late results) to the data are permitted.

NOTE: † The factor comes from the "standard" normal distribution, which has a mean of zero and a standard deviation (SD) equal to one. The interquartile range of such a distribution is $[-0.6745, +0.6745]$ and this is narrower than the familiar ± 1 SD interval. So, to convert an IQR into a ± 1 SD range, it must be scaled up by the ratio of the interval widths, namely $2/1.3490$. To then convert this ± 1 SD range (whose width is 2 standard deviations) into an amount equivalent to 1 SD, this range is then halved. Hence the IQR is divided by 1.3490 (or equivalently multiplied by 0.7413) to convert it into an estimate of the standard deviation.

A.5 Robust Z-scores & Outliers

KAN uses z-scores based on robust summary statistics (the median and normalised IQR) to statistically evaluate the participants' results. Where pairs of results have been obtained (i.e. in most cases), two z-scores are 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 called A and B. The median and normalised IQR of all the sample A results are denoted by median(A) and normIQR(A), respectively. (Similarly for sample B.) A simple robust z-score (denoted by Z) for a laboratory's sample A result would then be:

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

The standardised sum (denoted by S) and standardised difference (D) for the pair of results are:

$$S = (A + B)/\sqrt{2} \text{ and } D = (B-A)/\sqrt{2} \text{ if median}(A) < \text{median}(B) \\ (A-B)/\sqrt{2} \text{ otherwise}$$

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. (these summary statistics are usually tabled in the report, to allow participants to calculate the z-scores themselves).

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.

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

The calculated z-scores are tabulated in the report for a program, alongside the corresponding results (see example on page 21) and the results are assessed based on their z-scores. An outlier is defined as any result/ pair of results with an absolute z-score greater than three, i.e. $Z > 3$ or $Z < -3$. Outliers are identified in the table by a marker (**§**) beside the z-score.

This outlier criteria, $|Z| > 3$, has a confidence level of about 99% (related to the normal distribution) - i.e. there is a less than 1% chance that the result(s) is a true member of the population and it is far more likely that there is a problem with this result/pair of results. Similarly a z-score cut-off of two has a confidence level of approximately 95%. Laboratories which have a z-score in this region (i.e. $2 < |Z| < 3$) are encouraged to "take a close look at" their results.

When interpreting results which have been identified as outliers, the sign of the z-score and the design of the program must be considered. For both uniform and split pairs a positive between-laboratories outlier (i.e. $ZB > 3$) indicates that both results for that pair are too high. Similarly a negative between-laboratories outlier (i.e. $ZB < -3$) indicates that the results are too low.

For uniform pairs, where the results are on identical samples, a within-laboratory outlier of either sign (i.e. $|ZW| > 3$) indicates that the difference between the results is too large. For split pairs, where the analyte is at different levels in the two samples, a positive within-laboratory outlier (i.e. $ZW > 3$) indicates that the difference between the two results is too large and a negative within-laboratory outlier (i.e. $ZW < -3$) indicates that the difference is too small or in the 'opposite direction' to the medians.

For situations where a program involves a single result on one sample (X) a simple robust z-score is calculated as $Z = \{X - \text{median}(X)\} / \text{normIQR}(X)$ and outliers are classified as above - values of X for which $|Z| > 3$. 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), but whether this is due to between-laboratories or within-laboratory variation, or both, is unknown.

A.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 most commonly used graph as the ordered z-score bar-chart of which as described in detail below.

Ordered Z-score Chart

One of these charts is generated for each type of z-score calculated. 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 and -3, so the outliers are clearly identifiable as the laboratories whose "bar" extends beyond these cut-off lines. The y-axis is usually limited to range from -5 to +5, so in some cases very large or small (negative) z-scores appear as extending beyond the limit of the chart. The advantages of these charts are that each laboratory is identified and the outliers are clearly indicated.

APPENDIX B EVALUATION PROCEDURES FOR CALIBRATION PROGRAMS

B.1 Introduction

B.2 En Ratio

B.3 Uncertainty of Measurement

B.4 Graphical Displays

B.1 Introduction

This appendix outlines the procedures KAN uses to evaluate the results of its *calibration* programs (refer to Appendix B for procedures applicable to *testing* programs). The procedures used by NATA are consistent with those used for international calibration programs run by EAL and APLAC.

B.2 En Ratio

As stated in Section 7.6, NATA uses the E_n ratio to evaluate each individual result from a laboratory. E_n stands for **E**rror **n**ormalised and is defined as:

"

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

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 ratio (also called E_n number) should be between -1 and +1 i.e. $|E_n| < 1$. (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 uncertainty of measurement 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 ratios would be expected. So when considering the significance of any results with $|E_n|$ marginally greater than 1, 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 table of results from a 10 g mass standard interlaboratory comparison, their corresponding reported uncertainties and E_n ratios are tabulated below. The results for laboratories 7 and 8 are considered unsatisfactory.

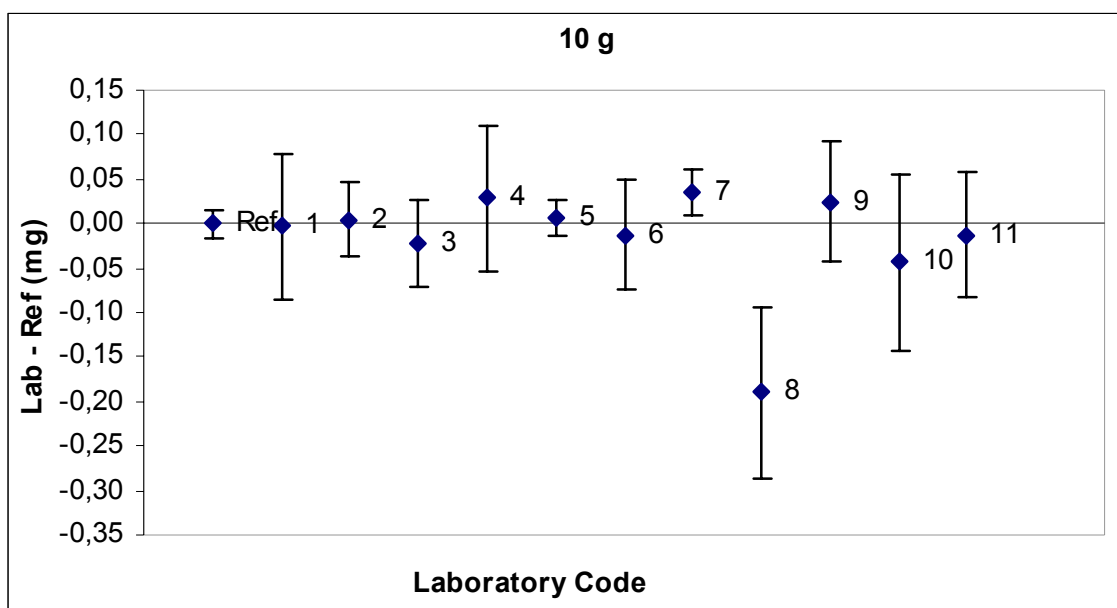
| Nominal mass 10 g | | | |
|-------------------|----------------|---------------|-------------|
| Lab.Code | Lab - Ref (mg) | U_{95} (mg) | E_n |
| Ref | 0 | 0,02 | |
| 1 | -0,004 | 0,08 | -0,04 |
| 2 | 0,00 | 0,04 | 0,1 |
| 3 | -0,02 | 0,05 | -0,4 |
| 4 | 0,03 | 0,08 | 0,3 |
| 5 | 0,01 | 0,02 | 0,3 |
| 6 | -0,01 | 0,06 | -0,2 |
| 7 | 0,04 | 0,03 | 1,2 |
| 8 | -0,19 | 0,10 | -1,9 |
| 9 | 0,02 | 0,07 | 0,4 |
| 10 | -0,04 | 0,10 | -0,4 |
| 11 | -0,01 | 0,07 | -0,2 |

B.3 Uncertainty of Measurement

The uncertainty of measurement reported by the laboratory is used in the E_n ratio. The artefacts used in these programs usually have sufficient resolution, repeatability and stability to allow the laboratory to report an uncertainty equal to their accredited "*least uncertainty of measurement*" as defined in their scope of accreditation (internationally referred to as their "*best measurement capability*"). If a laboratory reports an uncertainty larger than their accredited uncertainty then they would generally be asked for an explanation.

B.4 Graphical Displays

Graphs of reported results and their associated uncertainties are included in final reports. The example graph below shows a plot of the results tabulated in Section C.2. These graphs display each participant's LAB-REF value, represented by a black diamond w. The bars extending above and below the LAB-REF value represent the laboratory's reported uncertainty.



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