Determination of EVALUATION CRITERIA & different procedures for performance evaluation of participants

S. Subramanian

4.4.4 Statistical design

4.4.4.1 Statistical designs shall be developed to meet the objectives of the scheme, based on the nature of the data (quantitative or qualitative, including ordinal and categorical), statistical assumptions, the nature of errors, and the expected number of results.

NOTE 1: Statistical design covers the process of planning, collection, analysis and reporting of the proficiency testing scheme data. Statistical designs are often based on stated objectives for the proficiency testing scheme, such as detection of certain types of errors with specified power or determination of assigned values with specified measurement uncertainty.

NOTE 2: Data analysis methods could vary from the very simple (e.g. descriptive statistics) to the complex, using statistical models with probabilistic assumptions or combinations of results for different proficiency test items.

4.4.4 Statistical design

NOTE 3: In cases where the proficiency testing scheme design is mandated by a specification given by, for example, a customer, regulatory authority or accreditation body, the statistical design and data analysis methods can be taken directly from the specification.

NOTE 4 : In the absence of reliable information needed to produce a statistical design, a preliminary interlaboratory comparison can be used.

4.4.4.2 The proficiency testing provider shall document the statistical design and data analysis methods to be used to identify the assigned value and evaluate participant results, and shall provide a description of the reasons for their selection and assumptions upon which they are based. The proficiency testing provider shall be able to demonstrate that statistical assumptions are reasonable and that statistical analyses are carried out in accordance with prescribed procedures.

4.4.4 Statistical design

4.4.4.3 In designing a statistical analysis, the proficiency testing provider shall give careful consideration to the following:

a) accuracy (trueness and precision) as well as the measurement uncertainty required or expected for each measurand or characteristic in the PT;

b) the minimum number of participants in the proficiency testing scheme needed to meet the objectives of the statistical design;

- in cases where there is an insufficient number of participants to meet these objectives or to produce statistically meaningful analysis of results, the proficiency testing provider shall document, and provide to participants, details of the alternative approaches used to assess participant performance;

4.4.4 Statistical design

PT provider has to give careful consideration to :

c) the relevance of significant figures to the reported result, including the number of decimal places;

d) the number of proficiency test items to be tested or measured and the number of repeat tests, calibrations or measurements to be conducted on each proficiency test item or for each determination;

e) the procedures used to establish - standard deviation for proficiency assessment SDPA or other evaluation criteria such as d_E;

4.4.4 Statistical design

PT provider has to give careful consideration to :

f) procedures to be used to identify or handle outliers, or both;

g) where relevant, the procedures for the evaluation of values excluded from statistical analysis; and

h) objectives to be met for the design of PT round

4.7 Data Analysis & evaluation of PT Scheme results
4.7.1 Data Analysis and records

4.7.1.1 All data processing equipment and software shall be validated in accordance with procedures before being brought into use. Computer system maintenance shall include a back-up process and system recovery plan. The results of such maintenance and operational checks shall be recorded.

4.7.1.2 Results received from participants shall be recorded and analysed by appropriate methods. Procedures shall be established and implemented to check the validity of data entry, data transfer, statistical analysis, and reporting.

4.7.1.3 Data analysis shall generate summary statistics and performance statistics, and associated information consistent with the statistical design of the proficiency testing scheme.

4.7 Data Analysis & evaluation of PT Scheme results
4.7.1 Data Analysis and records

4.7.1.4 The influence of outliers on summary statistics shall be minimized by the use of robust statistical methods or appropriate tests to detect statistical outliers.

4.7.1.5 The proficiency testing provider shall have documented criteria and procedures for dealing with test results that may be inappropriate for statistical evaluation, e.g. miscalculations, transpositions and other gross errors.

4.7.1.6 The proficiency testing provider shall have documented criteria and procedures to identify and manage proficiency test items that have been distributed and are subsequently found to be unsuitable for performance evaluation, e.g. because of in-homogeneity, instability, damage or contamination.

4.7 Data Analysis & evaluation of PT Scheme results
4.7.2 Evaluation of Performance

4.7.2.1 The proficiency testing provider shall use **valid methods of evaluation** which meet the purpose of the proficiency testing scheme. The methods shall be documented and include a description of the basis for the evaluation. **The evaluation of performance shall not be subcontracted** (see 5.5.2).

4.7.2.2 Where appropriate for the purpose of the proficiency testing scheme, the proficiency testing provider shall provide **expert commentary on the performance of participants with regard to the following**

4.7 Data Analysis & evaluation of PT Scheme results
4.7.2 Evaluation of Performance

The expert comments concern:

a) overall performance against prior expectations, taking measurement uncertainties into account;

b) variation within and between participants, and comparisons with any previous proficiency testing rounds, similar proficiency testing schemes, or published precision data;

4.7 Data Analysis & evaluation of PT Scheme results
4.7.2 Evaluation of Performance

The expert comments concern:

c) variation between methods or procedures;

d) possible sources of error (with reference to outliers) and suggestions for improving performance;

e) advice and educational feedback to participants as part of the continual improvement procedures of participants;

4.7 Data Analysis & evaluation of PT Scheme results
4.7.2 Evaluation of Performance

The expert comments concern:

f) situations where unusual factors make evaluation of results and commentary on performance impossible;

g) any other suggestions, recommendations or general comments; and

h) conclusions.

Determination of EVALUATION CRITERIA

EVALUATION CRITERIA

- **Standard Deviation for Proficiency Assessment (SDPA)- Measure of dispersion used in the evaluation of results of Proficiency Testing**
- **Maximum permissible measurement error (Limit of Error) Delta E (ðE)- Extreme value of measurement error, with respect to a known reference quantity value, permitted by specifications or regulations for a given measurement, measuring instrument, or measuring system**

ISO/IEC 17043 - Annex B.3 – Calculation of SDPA (II)

1. By perception of experts (8.2)
2. By experience from previous rounds of a PT Scheme (8.3)
3. By use of a General model (8.4)
4. Using the repeatability and reproducibility standard deviations from a previous collaborative study of precision of a measurement method (8.5)
5. From data obtained in the same round of a PT Scheme (8.6)

SDPA – By perception of experts (8.2)

- SDPA may be set at a value required for a specific task of data interpretation
- Determined by expert judgement or regular mandate or Accreditation Body
- From "outside of the PT system"
  E.g.: Directive (2009/90/EC) from the European Commission to the Water Framework Directive WFD: Analytical method can have a max. uncertainty of 25 %
- E.g Analytical method for aflatoxin in nuts not to have reproducibility CV of more than 50%

SDPA – By perception of experts (8.2)

- If a regulatory requirement or fitness for purpose goal is given as a SD, it may be directly used as SDPA
- If fitness for purpose goal is given as maximum permissible measurement error, it may be directly used as Delta E (ðE)
- Delta E (ðE) may be transformed to SDPA, by dividing this limit by the number of multiples of SDPA that are used to define an action signal (or unacceptable result) and vice versa

SDPA – By perception of experts (8.2)

- ✓ Choice according to a "fitness-for-purpose"-wish for the laboratory
- ✓ Fitness of purpose – decided by Regulatory authority, Accreditation Body or Technical experts of PT Provider
- ✓ This value can be unrealistic to the reproducibility of the method

SDPA – By perception of experts (8.2)

- PT Provider is to ensure that SDPA or other evaluation criteria is set at a reasonable value which is achievable by the participants
- For this calculate Between Lab SD , Sigma L from Reproducibilty and Repeatability SD of the measurement method
- Calculate Ø using the following formula
  SDPA = SQRT[(ØxSigma L)^2 + Sigma r^2/n)
- Ø should be more than 0.5 to be realistic

By experience from previous rounds of a PT Scheme (8.3)

- ✓ Consider the assigned calue and SDPA fixed in a no. of accredited PT Schemes conducted by different PT Providers – for same matrix, measurand and concentration range
- ✓ Calculate SDPA %
- ✓ Fix the average or median of SDPA % for the current round.

SDPA – by use of a General model (8.4)

- Value of the SDPA is derived from a general model for the reproducibility of the measurement method (z.B. Horwitz curve)
  - Reproduciblity SD is a function from the concentration (mass fraction)
  - Has advantage of (a) objectivity (b) consistency across measurands & (c) being empirically based
- Disadvantage:
  - The true reproduciblity can vary from the value of the model
  - Reproducibility is dependent on the parameter, the method, sample and not only on the concentration

SDPA – Using the repeatability and reproducibility standard deviations from a previous collaborative study of precision of a measurement method (8.5)

- When one standardised method is used in the PT
- Requirement: Information of the repeatability and reproducibilty must be available
- Calculation of the SDPA using this information

SDPA – from data obtained in the same round of a proficiency testing scheme (8.6)

- Calculated with robust statistic from the results of the participants in PT
- ISO/TS 20612:
  - Q-Method
  - Application of a variance function
- ISO 13528
  - Algorithm A

IF CONSENSUS VALUE IS USED FOR DETERMINATION OF SDPA (8.6), THERE SHOULD BE ENOUGH PARTICIPANTS TO

ACHIEVE SUFFICIENTLY LOW UNCERTAINTY TO REDUCE THE INFLUENCE OF OUTLIERS

SDPA – from data obtained in the same round of a proficiency testing scheme (8.6)

Advantages:

This approach has simplicity and conventional acceptance due to successful use

Disadvantages:

- **SDPA may very from round to round and Trend analysis is difficult**
- **SDPA can be unreliable when (a) "p" is small and (b) results from different methods are combined**
- **Using dispersion measures from the data will lead to constant proportion of "APPARENTLY ACCEPTABLE " scores**

Performance evaluation of participants

DIFFERENT PERFORMANCE STATISTICS

1. **ESTIMATE OF DEVIATION – D & D% (9.3)**
2. **z – Score (9.4)**
3. **z' - Score (9.5)**
4. **Zeta Score (ζ) (9.6)**
5. **En Score (9.7)**
6. **PA Score (9.8)**

EVALUATION CRITERIA – GENERAL CONSIDERATIONS

- **Statistics used for performance evaluation – to be consistent with OBJECTIVES of PT Scheme**
- **Performance statistics are meaningful only when participants results are reviewed & are determined to be consistent with the DESIGN of PT Scheme. i.e No evidence of**
  - **deterioration of PT items**
  - **mixture of populations (groups) of participants**
  - **no severe violation of any statistical assumptions of PT data**
- **Not appropriate to use evaluation methods that intentionally classify a fixed proportion of results as generating an "Action Signal"**

Universität Stuttgart

Calculation of performance statistics – percent difference $D_{\%}$

- Estimate of laboratory bias: $D = (x-X)$
- x: participant's result
- X: assigned value
- Percent difference $D_{\%}$: $D_{\%} = \frac{(x-X)}{X} * 100$
- Independent of the magnitude of the assigned value

Universität Stuttgart

Calculation of performance statistics – z-Score

$$z = \frac{(x-X)}{\hat{\sigma}}$$

x result of the participant
X assigned value
$\hat{\sigma}$ SDPA

- Adoption: data set is normal distributed

Calculation of performance statistics – z'-Score

$$z' = \frac{(x - X)}{\sqrt{\hat{\sigma}^2 + u_x^2}}$$

u_x standard uncertainty of the assigned value

- Application of z-Scores, if:

$$u_x \leq 0{,}3\hat{\sigma}$$

- Otherwise the uncertainty of the assigned value is not negligible
- Then the possibility is given that z-values deliver a warning or action signal, but not the z'-values

Calculation of performance statistics – zeta-Score

$$\zeta = \frac{x - X}{\sqrt{u_x^2 + u_X^2}}$$

- u_x estimate of the standard uncertainty from the result of the laboratory
- Application and interpreation of zeta-Scores
  - ζ-Scores can be used together with z-scores to check the plausiblity of the estimation of measurement uncertainty

Calculation of performance statistics – E_n-Score

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{ref}^2}}$$

- X assigned value derived from a reference laboratory
- U_{ref} *expanded uncertainty from X*
- U_{lab} expanded uncertainty from the result x of a laboratory
- Applied in key comparisons of metrology institutes

IF En SCORE OR ZETA SCORE IS USED, THESE WILL BE MEANINGFUL ONLY IF THE UNCERTAINTY ESTIMATES ARE DETERMINED BY THE PARTICIPANTS IN A CONSISTENT MANNER USING ISO/IEC GUIDE 98

Universität Stuttgart

Calculation of performance statistics

z-, z_U-, z'-, zeta-score:

$|z| \leq 2,0 \Rightarrow$ satisfactory

$2,0 < |z| < 3,0 \Rightarrow$ questionnable

$|z| \geq 3,0 \Rightarrow$ unsatisfactory

E_n-Score:

$|E_n| \leq 1,0 \Rightarrow$ satisfactory

$|E_n| > 1,0 \Rightarrow$ unsatisfactory

REPORTING OF PERFORMANCE EVALUATION

- **For small data (generally with less than 8/12/18 results) there will be increased uncertainty of Assigned Value (if it is determined by "Consensus Value)**
- **Hence, performance score may be given for an idea only that too with a remark about the high uncertainty of Assigned Value which could have influenced the evaluation**

REPORTING OF PERFORMANCE EVALUATION

- In case of "wrong" results reported by participants – due to calculation error or use of incorrect units etc., it may be difficult for PT Provider to identify the same as **"WRONG"**
- Hence, **initially** Robust average and SDPA may be calculated taking in to account all results (excluding such results which are reported as < 25 mg/kg or > 25 mg/kg) .
- Those results which are outside the range of Assigned Value **± 5 times SDPA** may be excluded and the Robust average and SDPA are re-calculated, which only will be used for final statistical analysis of participants results.
- All results, including excluded results (but excluding such results which are reported as < 25 mg/kg or > 25 mg/kg) are given performance score.

INTERPRETATION OF PARTICIPANT UNCERTAINTIES IN TESTING

- **Use of M.U calculated by participants for performance evaluation is common in PT Schemes of calibration and not in PT Schemes of testing**
- **M.U reported by testing labs can be used**
  - **AB can examine whether MU reported are consistent with their scope of accreditation**
  - **Participants can review consistency of MU reported by comparing with other labs**
  - **Can be used for confirmation of claims (VALIDATION)of Uncertainty**

SCREENING OF UNCERTAINTIES REPORTED BY PARTICIPANTS

- **When $u(x_{pt}) \leq 0.3$ SDPA, it is unlikely that a participants std. uncertainty < $u(x_{pt})$. Hence, $u(x_{pt})$ can be used as lower limit (u_{min}) for screening of std. uncertainties reported by participants in a PT Scheme.**
- **Similarly, it is unlikely that a participants std. uncertainty > (1.5 x Robust SD of participant results) Hence, this can be used as upper limit (u_{max}) for screening of std. uncertainties reported by participants in a PT Scheme.**