ISO 13528:2015
Statistical methods for use in proficiency testing by interlaboratory comparison

ema training workshop

August 8-9, 2016

Mexico City

Class Schedule

- Monday, 8 August
- Types of PT of interest
  - ISO 17043?
  - Application of 13528
- Introduction to ISO 13528
  - Requirements in 17043 for statistical methods
  - Objectives for the scheme
    - Agree with peers?

Class Schedule

- Monday
- Requirements in ISO 13528
- Estimating the mean and SD
  - Median, nIQR
  - Algorithm A
  - Hampel/Qn
- Visual review of data
  - Histogram, kernel density
- Exercises: Algorithm A, kernel, histogram

Class Schedule

- Tuesday 9 August
- Homogeneity and Stability
  - Use of Experience
  - Classical design
- Scores: $z \quad z' \quad zeta \quad E_n \quad D$
- Other Design Issues
- Exercises: Homogeneity, Stability

ISO 13528:2005 - Development

- Written by ISO TC69, SC6
- Published in 2005, reaffirmed in 2009
- Some parts are widely used:
  - – Robust Algorithm A for mean and SD
  - – Homogeneity and stability procedures
  - – Uncertainty of assigned values
  - – Scores: z, z', $zeta$, E_n, D, $D\%$
- Revised 2010-2015

ISO 13528:2015 - Development

- ILAC Work Item proposal, June, 2010
- Published August, 2015
  - Minor corrections, June 2016
- Adopted by many PT providers and accreditation bodies as guidance for application of ISO/IEC 17043
  - Mandatory compliance by some bodies

Requirements for Statistical Methods – ISO/IEC 17043

- **4.4.1.3** Design of proficiency testing scheme
  - p) detailed description of the statistical analysis to be used;
  - q) the origin, metrological traceability and measurement uncertainty of any assigned values;

- **4.4.1.4** PTP shall have access to the necessary technical expertise and experience in the relevant fields of testing, …as well as statistics

Requirements for Statistical Methods – ISO/IEC 17043

- 4.4.3.1 Criteria for suitable homogeneity and stability shall be established and shall be based on the effect that inhomogeneity and instability will have on the evaluation of the participants' performance.
- 4.4.3.2 The procedures for the assessment of homogeneity and stability shall be documented and conducted in accordance with appropriate statistical designs.

Requirements for Statistical Methods – 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.

Requirements for Statistical Methods – Statistical Design

- 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 PTP shall be able to demonstrate that statistical assumptions are reasonable and

  - that statistical analyses are carried out in accordance with prescribed procedures.

Requirements for Statistical Methods – Assigned Values

- 4.4.5.1 The proficiency testing provider shall document the procedure for determining the assigned values…. This procedure shall take into account the metrological traceability and measurement uncertainty required to demonstrate that the proficiency testing scheme is *fit for its purpose*.

- 4.4.5.4 When a consensus value is used as the assigned value, the proficiency testing provider shall document the reason for that selection and shall estimate the uncertainty of the assigned value as described in the plan for the proficiency testing scheme.

Requirements for Statistical Methods – Data Analysis

- 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.
- 4.7.1.4 The influence of outliers on summary statistics shall be minimized by the use of robust statistical methods or appropriate tests to detect statistical outliers.
- 4.7.2.1 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.

QUESTIONS

Big Change in ISO 13528:2015

- Implement requirements of ISO/IEC 17043
  - – Assume adequate experience and statistical expertise
  - – State objectives for the scheme
    - What is the purpose? (to know “fit for purpose”)
  - – Use statistical methods that are appropriate to meet the objectives of the scheme
  - – Understand statistical assumptions and demonstrate they are reasonable
  - – NO BLIND APPLICATION OF FORMULAE

Introduction to ISO 13528

- Sections 0.1-0.5
  - Purposes of proficiency testing
    - Same as ISO/IEC 17043 Introduction
  - Rationale for scoring
    - Use participant results or independent criteria
  - ISO 13528 and ISO/IEC 17043
    - New sections in 13528, new topics in ISO 17043
  - Statistical expertise – Read this
  - Computer software
    - Must be validated

Section 1: Scope

- For Providers of PT schemes:
  - Detailed descriptions of statistical methods for design of scheme, and
  - For analysis of data from the PT scheme
- For participants and accreditation bodies:
  - Statistical methods to interpret PT data.
- General:
  - Can be used to demonstrate acceptable performance relative to specific criteria
  - Procedures for quantitative & qualitative data
  - Can be applied in Inspection

Section 2: Normative References

- Documents “indispensible” for application
  - ISO/IEC 17043
  - ISO Guide 30: Terms and definitions for RM
  - ISO 3534 : Statistics vocabulary and symbols
  - ISO 5725: Accuracy of measurement methods and results
  - ISO Guide 99: VIM

3: Terms and Definitions

- Most terms and definitions from normative references, some repeated for clarity
  - PT, PT item, PT provider, PT scheme, ILC, participant, measurement error
- Some terms modified slightly
  - SDPA, assigned value, outlier, PT item
- Some new definitions for this document
  - Consensus value, action signal

4. General principles

- 4.1 General requirements

4.1.1 The statistical methods used shall be fit for purpose and statistically valid. Any statistical assumptions on which the methods or design are based shall be stated in the design or in a written description of the proficiency testing scheme, and these assumptions shall be demonstrated to be reasonable.

NOTE on what is “statistically valid”

4. General principles

- 4.1 General requirements

  4.1.2 The statistical design and data analysis techniques shall be consistent with the stated objectives for the PT scheme.

  4.1.3 The PT provider shall provide participants with a description of the calculation methods used, an explanation of the general interpretation of results, and a statement of any limitations relating to interpretation.

4. General principles

- 4.2 Basic Model

4.2.1 For quantitative results in proficiency testing schemes where a single result is reported for a given proficiency test item:

$$x_i = \mu + \varepsilon_i$$

With x_i = proficiency test result for participant i

μ = true value for the measurand

ε_i = measurement error for participant i

4. General principles

- NOTE 1 Common models for ε include: the normal distribution $\varepsilon_i \sim N(0, \sigma^2)$ with mean 0 and variance either constant or different for each laboratory; or more commonly, an ‘outlier-contaminated normal’ distribution consisting of a mixture of a normal distribution with a wider distribution representing the population of erroneous results.

4. General principles

- NOTE 2 The basis of performance evaluation with *z* scores and σ_{pt} is that in an “idealized” population of competent laboratories, the interlaboratory standard deviation would be σ_{pt} or less.
- NOTE 3 This model differs from the basic model in ISO 5725, in that it does not include the laboratory bias term B_i. This is because the laboratory bias and residual error terms cannot be distinguished when only one observation is reported.

4.3 General approaches for the evaluation of performance

4.3.1 There are three different general approaches for evaluating performance in a proficiency testing scheme. These approaches are used to meet different purposes for the proficiency testing scheme.

a) performance evaluated by comparison with externally derived criteria;

b) performance evaluated by comparison with other participants;

c) performance evaluated by comparison with claimed measurement uncertainty.

4.3 General approaches for the evaluation of performance

- **4.3.2** The general approaches can be applied differently for determining the assigned value and for determining the criteria for performance evaluation.
- [for example, reference mean and consensus σ_{pt} ; or consensus mean and reference σ_{pt}.]
- In approach c) using measurement uncertainty, the assigned value is typically an appropriate reference value *(difficult to do with a consensus assigned value).*

5. Statistical design of proficiency testing schemes

- *This section presents a summary of the main points of ISO 13528:2005, adapted for the new Basic Model and for meeting the objectives for the PT scheme.*
  - *ISO 13528:2005 seeks to provide a good estimate of laboratory bias* B_i
  - *This revision seeks to evaluate the fitness of the PT result* x_i

5. Statistical design of proficiency testing schemes

5.1 Introduction to the statistical design of proficiency testing schemes

5.2 Basis of a statistical design

5.3 Considerations for the statistical distribution of results

5.4 Considerations for small numbers of participants

5.5 Guidelines for choosing the reporting format

5.1 Introduction to the statistical design

- PT does not generally evaluate lab bias or precision (but could if that is an objective)
- Evaluates fitness of a result as it would be submitted to a customer
  - Based on difference from the best estimate of "correct"
- Examination over several rounds can indicate bias and poor precision

5.2 Basis of a statistical design

- Design must be appropriate for the stated objective**s** for the scheme
- Quantitative or qualitative data
  - Quantitative: interval or ratio scale
  - Nominal / Ordinal scale
- Statistical assumptions
- Nature of errors
- Expected number of results

5.3 Considerations for statistical distribution

- Most techniques assume normal distribution for results from competent labs
  - Usually contaminated (bias or imprecision)
  - No need to verify normality
  - Check for symmetric and unimodal
- Transform data if necessary
- Use other appropriate distribution if needed
- State the basis for the design

5.4 Considerations for small numbers

- ISO/IEC 17043 requires consideration of what to do with fewer results than expected
  - – IUPAC/CITAC says to use CRMs
- Minimum number depends on several factors
  - – “Its not the size that counts…”
- See Annex D for further guidance on small numbers of results

5.5 Guidelines for report format

- Provider could ask specified format, but should request results are generated and reported the same as for customers
- If replicate results are requested, record all
  - Not just the mean or SD
- Have design consideration for “<“ (and “>”)
- Rounding error should be negligible
- If participants can report different formats, need to take that into consideration

6 Initial review of proficiency testing items and results

- 6.1 Homogeneity and stability of PT items
  - – References Annex B (normative)
  - – Options for experienced PT schemes
- 6.2 Different measurement methods
- 6.3 Blunder removal
- 6.4 Visual review of data
- 6.5 Robust statistical methods
  - – References Annex C
- 6.6 Outlier techniques for individual results

6.1 Homogeneity and stability of PT items

- Three alternatives offered:
  - Experimental studies as in Annex B
  - Use of experience on “closely similar” items
  - Assess participant results, compare SD
- Calibration: assure stability throughout
- Usually check all measurands or have defined correlation between tested and not
  - Understand what could cause inhomogeneity

6.2 Different measurement methods

- Should normally have the same assigned value for all methods that have the same measurand
  - – Not always possible (e.g., IVDD (medical))
- Need for same or different assigned values must be considered in the design
  - – *Design could allow flexibility*

6.3 Blunders

- Remove blunders prior to data analysis
  - Based on technical judgment and experience
    - You should know it when you see it
  - Can affect robust techniques and outlier detection routines

- When in doubt, do not discard
  - Robust techniques will minimize the effect

6.4 Visual review

- Expect unimodal and symmetric for most techniques
- Look for bimodal, asymmetric, or a large set of statistical outliers (minor modes)
  - Histogram
  - Kernal density plot
- Might have different procedure for first-time PT than for well established schemes

6.5 Robust techniques

- Robust techniques preferred to outlier removal. Better to retain all results that were not obvious blunders.
- Most techniques base estimates on the center 50% of the distribution
  - Median and nIQR or MADe
  - Algorithm A (and Algorithm S for precision)
  - Q/Hampel

6.6 Outlier techniques

- Can be useful to support visual review for blunders, but not optimal for extreme values
  - Assumptions underlying the test must be demonstrated to be appropriate
- Rejection strategies are allowed when robust methods are not applicable.
- If a result is removed, it should be evaluated according to criteria used for all participants.

7 The assigned value and its standard uncertainty

- 7.1 Choices for determining the assigned value
  - – *Five alternatives are discussed*
- 7.2 The uncertainty of the assigned value
  - – *A measurement is incomplete without its uncertainty*
- 7.3-7.7 *Different approaches that are allowed*
- 7.8 Comparison of the assigned value with a reference value
  - – *A consensus value might be biased*
  - – *A reference value might be unachievable*

7.1 Choice of method of determining the assigned value

- Alternative methods may be used if they have a sound statistical basis and the method is described in the plan for the scheme.
  - – Regardless of the method chosen, it must be checked for every round
- The method used must be fully described to participants in every report (or referenced)

7.2 Determining the uncertainty of the assigned value

- Reference to GUM and ISO Guide 35
- As in Guide 35:

$$u(x_{pt}) = \sqrt{u_{char}^2 + u_{hom}^2 + u_{trans}^2 + u_{stab}^2}$$

- Some components can reasonably be expected to be zero, based on experience.
- Concern that bias in assigned value is not accounted for.

7.3 Formulation

- Based on preparation with known materials and calculation of properties
- Concern about representativeness of formulation versus naturally incurred
- This is a general approach which also applies when a reference value is determined by a primary method (see ISO Guide 35)
- Standard uncertainty for characterization is determined by an appropriate model for formulation or a primary method.

7.4 Certified reference material

- Assigned value from certified property value
  - Concern if CRM is known to participants
  - Concern if reference material is representative of natural materials
- Standard uncertainty of the certified assigned value includes homogeneity and stability components.

7.6 Consensus value from expert laboratories

- Using a design for an interlaboratory study for characterization, as described in ISO Guide 35
  - Each participant must provide their uncertainty
  - PTP must have a procedure to combine uncertainties (no consensus in ISO or REMCO on this)
- If experts provide single results and no uncertainty, follow procedures in clause 7.7
- If experts provide multiple replicate values and no uncertainty, PTP must have a design
  - This also applies if there is evidence that some uncertainties are not correctly determined

7.7 Consensus value from participants

- Use techniques described in Annex C
  - Careful application of techniques in clauses 6.2-6.6 to assure that adequate agreement exists and assumptions are demonstrated to be reasonable
  - May wish to use a subset of participants
  - Can use other calculation methods
- Some advantages
  - No additional measurements needed
  - May be necessary with operationally-defined measurands

7.7 Consensus value from participants

- Many disadvantages
  - There may be insufficient agreement
  - Consensus value can be biased due to faulty methods or biased methods
    - This can lead to underestimate of uncertainty
  - No metrological traceability
- Uncertainty of characterization from the method used. For some robust methods:

$$\mu(x_{pt}) = 1.25 \times s^{*} / \sqrt{p}$$

7.8 Compare assigned value with independent reference value

- When consensus value is used as x_{pt}, then PTP should obtain independent reference value (formulation, expert, etc). Called x_{ref}
- When reference value is x_{pt}, then should compare with consensus mean
- Calculate $x_{diff} = (x_{ref} – x_{pt})$

$$u_{diff} = \sqrt{u^2(x_{ref}) + u^2(x_{pt})}$$

Criterion for acceptance: $| x_{diff} | < 2u_{diff}$

8. Determination of criteria for evaluation of performance

- 8.2 By perception of experts
- 8.3 By experience from previous rounds of a proficiency testing scheme
- 8.4 By use of a general model
- 8.5 Repeatability and reproducibility SD from a collaborative study of precision
- 8.6 From data obtained in the same round of a proficiency testing scheme

8.1 General approaches

- Basic approach is to compare participant result with x_{pt} and compare the difference with an allowance for measurement error
  - Using a standardized performance statistic
    - z score, z' score, zeta score, E_n,
  - Difference might be defined (e.g., regulation)
    - D and $D\%$, or “within limits” / ”not within limits”
- Can be useful to have standardized scope to compare across rounds

8.2 Perception of experts

- Allowance for error can be determined by technical experts, accreditation bodies, or regulatory bodies.
  - Can be expressed as Standard Deviation for Proficiency Assessment (SDPA) : σ_{pt}
  - Can be expressed as Maximum Permissible Error: δ
- If criterion for acceptable performance is z<3.0, then $\delta = 3\sigma_{pt}$ and $\sigma_{pt} = \delta/3$

8.3 Experience with previous rounds of PT

- When a PTP has experience over several rounds with similar PT items, measurands, and methods, then σ_{pt} can be anticipated (see Annex E.8)
- Several advantages:
  - Evaluations based on reasonable criteria
  - Criteria will not vary from round to round due to random error or changing participant base
  - Criteria will not vary by PT provider
- Previous round data need to be checked for consistency and perhaps for performance by competent participants (not all participants)

8.4 Use of a general model

- Can use a general model for reproducibility σ_R to be used as σ_{pt}
- Results must be reasonable (don't use blindly as a default)
- Only one example is described (modified Horwitz Curve), but others might be possible.

With c = mass fraction of measurand and $0 \le c \le 1$:

$\sigma_R = 0.22c$ when $c < 1.2 \times 10^{-7}$

$\sigma_R = 0.2c^{0.8495}$ when $1.2 \times 10^{-7} \le c \le 0.138$

$\sigma_R = 0.1c^{0.5}$ when $c > 0.138$

8.5 Use σ_r and σ_R from previous collaborative precision study

- If a previous collaborative study followed principles of ISO 5725-2, repeatability σ_r and reproducibility σ_R estimates can be used to determine σ_{pt}
- With m the number of replicate values:

$$\sigma_{pt} = \sqrt{\sigma_R^2 - \sigma_r^2(1 - 1/m)}$$

- When $m = 1$ then $\sigma_{pt} = \sigma_R$

8.6 From data obtained in the same round of PT

- Consensus SD can be used as σ_{pt}
  - Should use robust technique from Annex C
- Caution about SD being inappropriate for evaluation of performance – can be too large or can be too small for “fitness for use”
  - Should have limits for smallest SD to be used
  - Should have limits for largest SD that can be used
  - Should have limits on range of values that can be evaluated as “acceptable”, based on fitness for use (for example, a minimum acceptable recovery of a formulated level)

8.6 From data obtained in the same round of PT

- Advantages of this approach
  - Easy, commonly used, may be only feasible approach
- Disadvantages
  - SD can vary widely from round to round
  - Can be unreliable with small number of labs
  - Can lead to approximately same proportion of “action signals” (unacceptable)
  - There is no useful interpretation of suitability of a result based on intended use (shows only that a lab agrees with others in the scheme). This can be important when the measurand involves health or safety.

8.7 Monitoring Interlaboratory agreement

- PTP should use a procedure to monitor interlaboratory agreement (robust SD) of participants across rounds
  - Useful for PTP to show benefits of participation
  - Useful to check suitability of statistical methods
  - Useful to check for unexpected increase or decrease in agreement

9. Calculation of performance statistics

- 9.1 General considerations
  - Statistics used for determining performance shall be consistent with the objectives for the PT scheme.
  - Results should be reviewed and determined to be consistent with the assumptions in the design
    - Approximate normality (unimodal, symmetric)
    - No signs of instability or inhomogeneity
    - Signs of mixed population

9.2 Limiting the uncertainty of the assigned value

- If the uncertainty of x_{pt} is large relative to the performance criterion, there is risk of adverse evaluations due to factors other than poor measurement technique

$$u(x_{pt}) < 0.3\sigma_{pt}$$

$$u(x_{pt}) < 0.1\delta_E$$

- This can be a difficult criterion. If it is exceeded:
  - Use a different assigned value
  - Accommodate the uncertainty in the evaluation (z', ζ, E_n)
  - Report different x_{pt} for different methods
  - Do not evaluate performance

9.3 Estimates of deviation (measurement error)

- All performance measures start with measurement error – deviation from the assigned value; expressed in units or %
- This deviation can be compared to a criterion, δ_E expressed in units or as a percentage of x_{pt} :

$D_i = (x_i - x_{pt}) < \delta_E$

$D_i \% = (x_i - x_{pt}) / x_{pt} < \delta_E$ if δ_E is a %

9.3 Estimates of deviation (measurement error)

- The error criterion, δ_E can be a regulatory limit, analytical goal, expert opinion, etc.
- A standardized score can be calculated, the percentage of allowed error, or P_A, expressed as a percentage:

$P_A = D_i / \delta_E \times 100\%$ or

$P_A = D_i \% / \delta_E$ if δ_E is a %

If $D_i > \delta_E$ then $P_A > 100\%$

9.4 z score

- The most commonly used statistic for PT
- In US medical applications, also called “standard deviation interval” (SDI)
- Calculated with SDPA σ_{pt} from clauses 8.2-8.6

$$z_i = \frac{(x_i - x_{pt})}{\sigma_{pt}}$$

- Generally: $|z| \leq 2.0$ ➔ “acceptable”

  $2.0 < |z| < 3.0$ ➔ “warning”

  $|z| \geq 3.0$ ➔ “unacceptable”

9.5 *z'* score

- A slight variation to z score, to allow consideration of uncertainty of x_{pt}
- When criterion in clause 9.2 is met

$$0.96 < z'/z < 1.00$$

- Calculated with SDPA σ_{pt} from clauses 8.2-8.6

$$z_i' = \frac{(x_i - x_{pt})}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}}$$

z' score is evaluated same as *z* score

9.6 Zeta score ζ

- If an objective of the scheme is to evaluate a result compared to the participant’s claim for uncertainty, can use

$$\text{Zeta} = \zeta = \frac{x_i - x_{pt}}{\sqrt{u^2(x_i) + u^2(x_{pt})}}$$

with $u(x_i)$ the standard uncertainty of result x_i

- Generally ζ can be interpreted the same as z score

9.7 E_n scores

- E_n (Error, normalized) is a conventional score for PT in calibration, but can be applied anywhere

$$E_n = \frac{x_i - x_{pt}}{\sqrt{U^2(x_i) \quad + U^2(x_{pt})}}$$

with $U(x_i)$ the expanded uncertainty of result x_i

- Generally | E_n | < 1.0 is “acceptable”

Caution: E_n and Zeta scores

- Scores that evaluate performance compared to claimed uncertainty must be interpreted with caution, because some participants might not calculate uncertainty correctly (GUM), or report them correctly.
- A large uncertainty leads to lower scores; small uncertainty leads to higher scores
- Often useful to report E_n and ζ in addition to a conventional score (*e.g.*, z z' D $D\%$)

9. 8 Evaluation of participant measurement uncertainties

- Proficiency testing is a useful tool for showing differences between laboratory measurements. This includes estimates of measurement uncertainty.
- Many laboratories and accreditation bodies could benefit from seeing that their estimates are much different than those of other laboratories using the same method
- ISO 13528 recommends informational ‘flags’ of questionable uncertainties.

9. 8 Evaluation of participant measurement uncertainties

- Reasonableness criteria for mu

$$u_{min} = u_{ref}$$

$$u_{max} = 1.5\sigma_{pt}$$

$u_{min} \leq u_{lab} \leq u_{max}$ ➔ OK

$u_{min} < u_{lab}$ ➔ u_{lab} may be small

$u_{lab} > u_{max}$ ➔ u_{lab} may be large

9.9 Combined performance scores

- Some PT schemes combine scores for different PT items in the same round (e.g., average z scores)
  - Useful when there are many samples or measurands
- Sometimes this is part of the design:
  - e.g., evaluation of precision or linearity
- Combined scores have unknown statistical properties, so should be used with caution
- Graphical techniques are preferred

10 Graphical techniques

- Graphs are encouraged, and are required in ISO/IEC 17043 (reports)
- Histograms are most common, for preliminary data checks and for reporting
  - – Kernal density plots are similar and easy
- Other techniques discussed in class

11 Design and analysis for qualitative schemes

- 11.1 Types of qualitative data
  - Nominal or categorical scale
  - Presence or absence (including above or below threshold)
  - Ordinal (response has magnitude, but no mathematical relationship between levels)

- Does NOT include
  - Count data
  - Quantitative results on a discrete scale (MPN)

11.2 Statistical design

- Homogeneity
  - Test suitable number of items
  - All results should be the same
- Stability
  - Should not be a factor in identity
  - Concern for presence if not stable

11.2 Statistical design

- Performance criterion based on expert judgment, often after review of results
  - Preferred to have a panel of experts, and defined criteria for their agreement
- Consider multiple samples or replicates:
  - Can have evaluation of detection levels, TP, TN, FP, FN

11.3 Assigned values for qualitative schemes

- Assigned value usually determined by expert opinion
  - – Categorical: can use participant mode
  - – Ordinal: can use participant median or mode
- Often need to document origin or source of PT item
- Uncertainty should not be a factor
  - – Exception: threshold and “indeterminate”

11.4 Performance evaluation

- Evaluation criteria must meet objectives and be fit for the purpose of the test.
- Criteria are usually determined by expert opinion
  - – Might be individual, based on expert review of each participant’s results
- Can have weighted performance score
    - “perfect” → score 0
    - “not perfect, but not bad” → score 1
    - “bad” → score 3

Annexes

- Informational annexes
  - Annex A: Symbols
  - Annex D: Additional guidance on statistical procedures
  - Annex E: Illustrative examples
- Normative annexes
  - Annex B: Homogeneity and stability of PT items
  - Annex C: Robust analysis

Annex B: Homogeneity and stability of PT items

- B.1 provides a basic design for a homogeneity experiment.
  - $g \geq 10$ samples
  - $m \geq 2$ replicates
    - Adjustment when m=1
  - Calculate SD between samples s_s
  - Check all measurands, unless correlated
- Criterion for repeatability of method

$$\sigma_r < 0.5\sigma_{pt}$$

Annex B: Homogeneity and stability of PT items

- B.2 provides acceptance criterion

$$s_s \leq 0.3\sigma_{pt} \quad \text{or} \quad s_s \leq 0.1\delta_E$$

- If criterion is not met:
  - – Include s_s in σ_{pt}
  - – Include s_s in $u(x_{pt})$ and use z' score
  - – When σ_{pt} is the robust SD of participant results, than sample differences are already included in σ_{pt}, so criterion “can be relaxed”

Annex B: B.4 Stability check

- If experience or technical reasons show stability can be expected for the time of PT study, then a limited stability study is adequate to show measurands were stable
  - – See ISO Guide 35 for more information
- Should check all measurands, unless …
- Two PT items are adequate if homogeneity is assured, else use >2 items
- Use more items or replicates if $\sigma_r > 0.5\sigma_{pt}$

Annex B: B.5 Stability criterion

- Simple experiment is to check mean of results on stability measurements ($\bar{y}_2$) versus mean of results from before shipment ($\bar{y}_1$ e.g., homogeneity check)
- Criterion for acceptance:

$$| \bar{y}_1 - \bar{y}_2 | \leq 0.3\sigma_{pt} \quad \text{or} \quad | \bar{y}_1 - \bar{y}_2 | \leq 0.1\delta_E$$

- If criterion is met, instability will not affect evaluations

Annex B: B.5 Stability criterion

- If criterion is not met:
  - – Consider if intermediate precision is source of difference | $\bar{y}_1$-$\bar{y}_2$| . If possible, use isochronus stability study, or a different method
  - – Increase $u(x_{pt})$ to include instability
  - – Expand criterion for acceptance
  - – Quantify the effect of instability and include it in the evaluation
  - – Examine production, storage, shipment to see if improvements are possible
  - – Do not evaluate performance

Annex B: B.6 Transport stability

- PT provider should check the effects of transport, at least initially (newly developed PT schemes).
  - Compare results on shipped items vs. stored items
  - Criterion for acceptance same as in B.5
  - Not required in ISO/IEC 17043, but required in ISO G34
- Any known effects should be considered in evaluation of performance, and included in $u(x_{pt})$
- If consensus mean and SD are used, then all samples may have same effect, so not an issue

Annex C: Robust analysis

- PT providers need to mitigate the effect of extreme results, because not all participants are competent, and extreme results are always possible. These results have a strong effect on consensus statistics
- There are two choices:
  - Remove statistical outliers
  - Use statistical techniques that are robust to these values
- Robust techniques are preferred

Annex C: Robust statistics

- Simple techniques:
  - – Median for x_{pt}
  - – $nIQR$ for σ_{pt}
  - – $MADe$ for σ_{pt}

- Conventional
  - – Algorithm A: for x^* and s^*
  - – Algorithm S: for s_r

Annex C: Robust statistics

- Computationally intense techniques:
  - Q_n for σ_{pt}
  - Q/Hampel for σ_{pt} and x_{pt}

- Useful when visual review is not possible or when a fully general technique is needed
  - Bimodal distributions should still be detected per sections 6.3 and 6.4

Annex D: Additional guidance

- Procedures for small numbers of participants
  - Items that could have been Notes in Clause 11
    - Identifying outliers
    - Estimates of location (mean)
    - Estimates of dispersion (SD)
- Efficiency and breakdown points for robust procedures
- Use of PT for evaluating σ_r and σ_R of a measurement method

Annex D.2: Compare estimators

- Breakdown point (bp): proportion of outliers that can affect the estimators:

Mean and SD:	bp > 0
Median:	bp ≥ 50%
MADe and nIQR	bp ≥ 50%
Algorithm A	bp ≥ 25%
Q/Hampel	bp ≥ 50%

- In general, bp ≥ 20% should be investigated

Annex D.2: Compare estimators

- Efficiency (e): ratio of the variance of an estimator compared with variance of mean and SD
  - Median: e ≈ 66%
  - MADe and nIQR e ≈ 37%
  - Algorithm A e ≈ 97% / 75%
  - Q/Hampel e ≈ 96% / 80%

- All estimators are unbiased

Annex D.3: Evaluate σ_r and σ_R

- Novel concept mentioned in ISO/IEC 17043, suggested in subsequent publictions
  - Based on assumptions of experience in PTP and participants
  - Estimates might be preferable to estimates from initial collaborative study
- Considerations are provided to assure that any subsequent estimates are reliable.
  - Number of participants
  - Repeatability of method
  - Multiple rounds
  - Consistency of data analysis procedures