Scoring systems for quantitative schemes – what are the different principles?

Dr.-Ing. Frank Baumeister
TGZ AQS-BW
at Institute for Sanitary Engineering, Water Quality and Solid Waste Management
of the University of Stuttgart
AQS Baden-Württemberg
Bandtäle 2
70569 Stuttgart
GERMANY
Tel.: +49 711 685 65442 / Fax: +49 711 685 55442
E-Mail: frank.baumeister@aqsbw.de
ISO/IEC 17043
4.7.2 Evaluation of performance

- 4.7.2.1 The proficiency 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.
ISO/IEC 17043 - Annex B.3 – Calculation of performance statistics

• Transformation of PT results into a performance statistic for interpretation and comparison purposes
• The purpose is to measure the deviation from the assigned value in a manner that allows comparison with performance criteria.
ISO/IEC 17043 - Annex B.2

Assigned value

1. Formulation
2. Certified reference value
3. Reference value
4. Consensus value from expert laboratories
5. Consensus value from participant’s data
ISO/IEC 17043 - Annex B.3 – Standard deviation for proficiency assessment

1. Prescribed value
2. By perception
3. Statistical model
4. From results of precision experiments
5. From participant results
SDPA – Prescribed value

- SDPA may be set at a value required for a specific task of data interpretation
- Determined by expert judgement or regular mandate
- From „outside of the PT system“
- E.g.: Directive (2009/90/EC) from the European Commission to the WFD:
  - Analytical method can have a max. uncertainty of 25 % at the EQS
SDPA – by perception

• Choice according to a „fitness-for-purpose“-wish for the laboratory
• Estimate from previous rounds of proficiency testing or expectations based on experience
• From „inside of the PT system“
SDPA – from a statistical model

• Value of the SDPA is derived from a general model for the reproducibility of the measurement method (e.g. Horwitz curve)
  • Reproducibility SD is a function from the concentration (mass fraction)
SDPA – from the results of a precision experiment

• When one standardised method is used in the PT
• Requirement:
  • Information of the repeatability and reproducibility must be available

Calculation of the SDPA using this information
SDPA – from data obtained in a round of a proficiency testing scheme

• 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
Performance scoring
Normal distribution – important properties


- The curve is symmetrical about the population mean μ
- The greater the value of σ the greater the spread of the curve
- Approximately 68% (68.27%) of the data lie within μ±1σ
- Approximately 95% (95.45%) of the data lie within μ±2σ
- Approximately 99.7% (99.73%) of the data lie within μ±3σ

\[
y = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}
\]
Estimate of laboratory bias $D$ – percent difference $D\%$

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

- **Warning signal if:** \( 2 \hat{\sigma} < D < -2 \hat{\sigma} \)
- **Action signal if:** \( 3 \hat{\sigma} \leq D \leq -3 \hat{\sigma} \)

- **Advantage:**
  - Easy to understand
- **Disadvantage:**
  - not standardised to allow simple scanning for action signals
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
Interpretation of z-scores

\[ |z| \leq 2,0 \Rightarrow \text{satisfactory} \]

\[ 2,0 < |z| < 3,0 \Rightarrow \text{questionnable (warning signal)} \]

\[ |z| \geq 3,0 \Rightarrow \text{unsatisfactory (action signal)} \]

- **Advantage:**
  - Standardised score
  - Most commonly used

- **Disadvantage:**
  - Performance assessment depends on the choice of SDPA
### $z'$-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
zeta-score ($\zeta$-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
- $u_X$ standard uncertainty from the assigned value

$\zeta$-score:
- $|\zeta| \leq 2,0 \Rightarrow$ satisfactory
- $2,0 < |\zeta| < 3,0 \Rightarrow$ questionable (warning signal)
- $|\zeta| \geq 3,0 \Rightarrow$ unsatisfactory (action signal)
Interpretation of $\zeta$-scores

- $\zeta$-scores can be used together with $z$-scores to check the plausibility of the estimation of measurement uncertainty.
**$E_n$-number**

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

- $X$ 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

**$E_n$-number:**

- $|E_n| \leq 1,0 \Rightarrow$ satisfactory
- $|E_n| > 1,0 \Rightarrow$ unsatisfactory
Laboratory assessment


• By combination of single value assessment
• Involves danger of misinterpretation
  • A laboratory can measure one parameter permanently wrong, but nevertheless is positive assessed
Combined Assessment According to IUPAC – RSZ


• RSZ (rescaled sum of z-scores)
  • RSZ = Σz/√m with m = number of scores
  • Same scale as z-score
  • Negative assessment, if all values are within the tolerance but a little biased in the same direction
  • Errors with opposite sign cancel each other out
Combined Assessment According to IUPAC – SSZ


- SSZ (sum of squared z-scores)
  - Does not consider the sign of z-scores
Control charts

- z-score
- upper action limit $z = 3$
- upper warning $z = 2$
- $z = 0$
- lower warning limit $z = -2$
- lower action limit $z = -3$

Baumeister: Scoring of EQA results, EQALM Symposium 2013, Bucharest