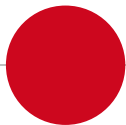


Performance assessment in non-quantitative PT

Marzia Mancin

on behalf of the Eurachem task force on non-quantitative PTs

10th Eurachem Workshop on Proficiency Testing in Analytical Chemistry, Microbiology and Laboratory Medicine
Windsor 25-28 September 2023

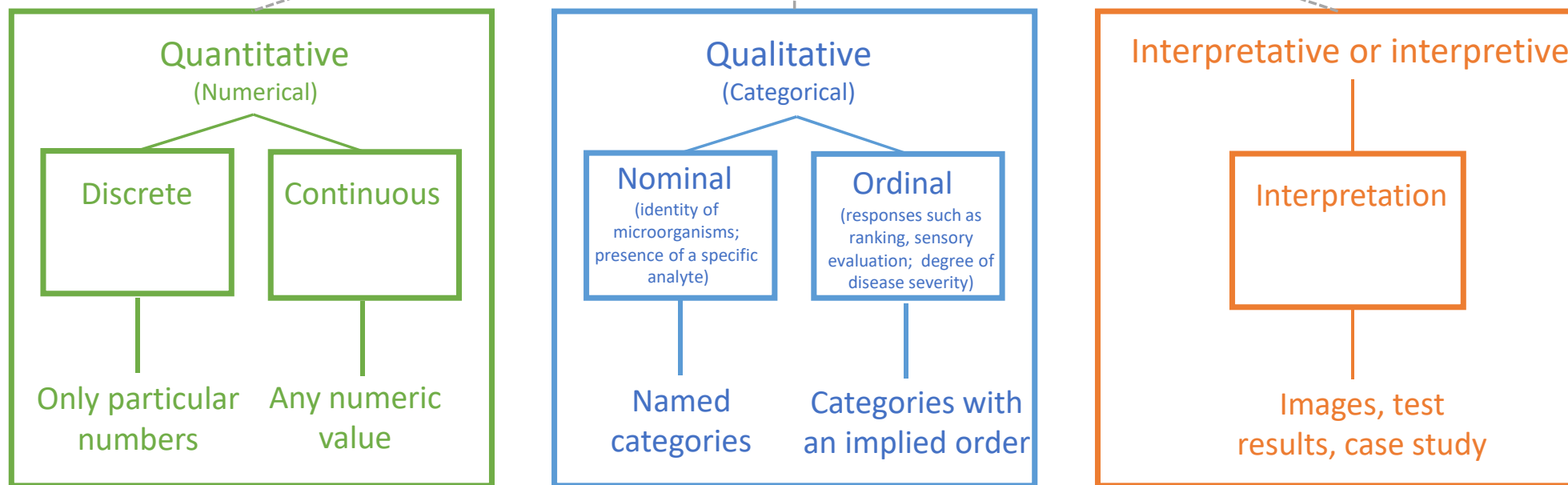


● Proficiency Testing

- Quantitative and non-quantitative Proficiency Testing (PT) or External Quality Assessment (EQA) have a relevant role in the evaluation of laboratory performance
- Laboratories use the PT performance evaluation to:
 - check the fair practice in obtaining results
 - monitor the results over time
 - verify the precision parameter
 - qualify the staff
- The PT provider has a great responsibility to provide a correct evaluation of the laboratory results

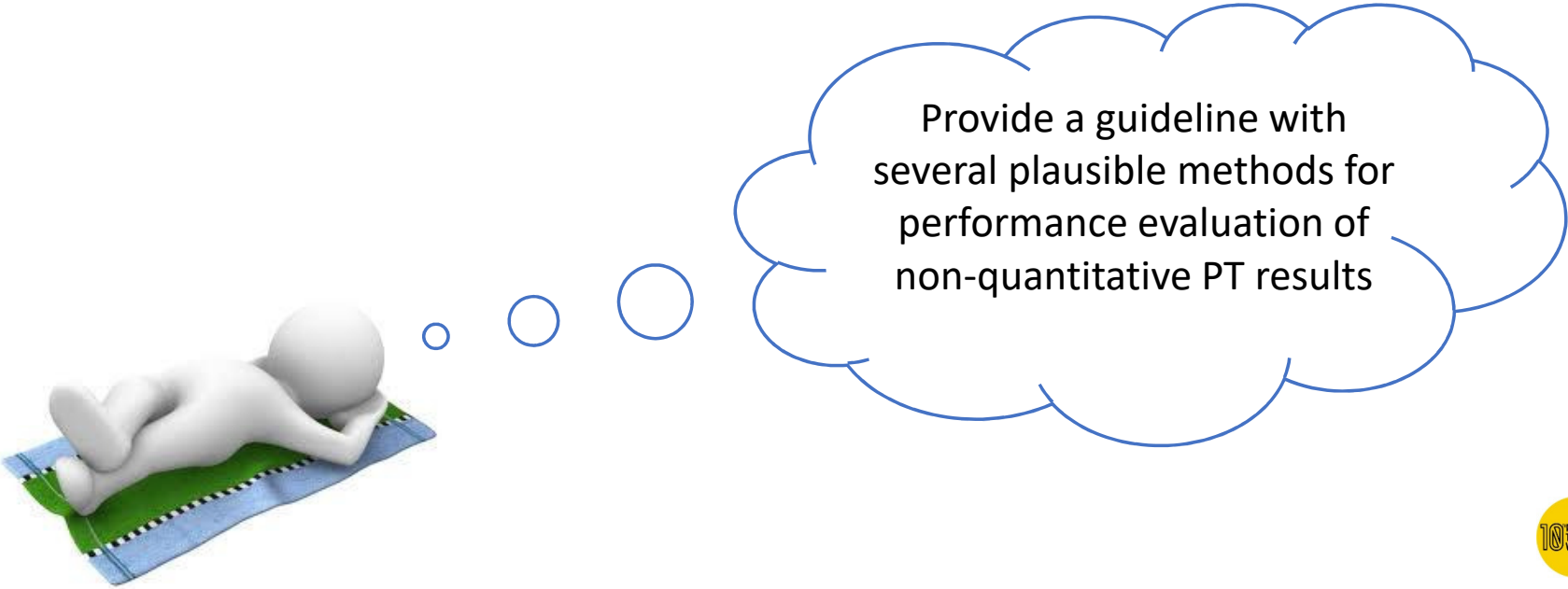
Types of Proficiency Testing

Type of results



● Why the focus on Non-quantitative PT performance evaluation?

- ISO standard 13528:2022, supporting ISO 17043:2023, describes in detail the statistics used for quantitative data
- On the contrary, guidelines for non-quantitative PT/EQA schemes assessment are limited
- Nevertheless, several statistical techniques are used by PT providers to evaluate performance in non-quantitative PT



Provide a guideline with several plausible methods for performance evaluation of non-quantitative PT results

● For this aim

- In 2014, the Eurachem PT working group realised an online survey with the aim to collect information on the performance evaluation of qualitative and interpretative PT/EQA



Tiikkainen, Ulla, et al. "*Is harmonisation of performance assessment in non - quantitative proficiency testing possible/necessary?*" *Accreditation and Quality Assurance* 27.1 (2022): 1-8. <https://doi.org/10.1007/s00769-021-01492-6>

- Most important results of the survey presented
- Literature review (up to 2020) on the used techniques for the performance evaluation of qualitative and interpretative PT/EQA

- I will present the global picture of non-quantitative PT performance assessment

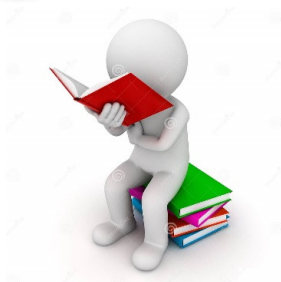
● Non-quantitative PT performance evaluation

- Tiikkainen et al. used three categories of scores for the performance evaluation:

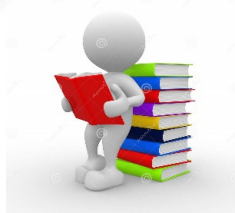
- Basic scores



- Elaborate scores



- Advanced scores



Basic scores

- Percentage of correct results (Snell, 1985 [1], and Vivienne, 2012 [2])
 - Attribution of a 1 or 0 score to correct or incorrect result, to assess the presence/absence or positivity/negativity of an item (e.g. antibiotics susceptibility testing)

For each sample, each laboratory has to identify the presence/absence of an antibiotic

Sample	Antib 1 Expected	Antib 1 Result	Score	Antib 2 Expected	Antib 2 Result	Score	Antib 3 Expected	Antib 3 Result	Score
1	+	+	1	+	+	1	+	-	0
2	+	+	1	-	-	1	+	-	0
3	-	-	1	+	+	1	+	-	0
4	-	-	1	-	-	1	+	+	1
.....
20	-	+	0	-	-	1	-	+	0
			90%			100%			75%

Incorrect result

Correct result

Basic scores

- **Score** (Vivienne, 2012 [2])

- Attribution of a four-point scoring to assess the “**degree of correctness**” of bacteriology or virology results
- Values such as “2, 1, 0 or -1” are assigned to «correct», «partially correct», «wrong», «grossly misleading» results
- Combination of scores and ranking them to provide the laboratories performance



Table 2 Example of weighted scoring applied to bacteriology specimens containing a single pathogen

Response	Core	Advanced
Correct species (with or without the correct serotype/toxin result)	2	2
Correct genus only	0	2
Correct species but incorrect serotype	0	1
Incorrect species	0	1
Negative result	0	0
Unnamed/unspecified microorganism	0	0
Incorrect genus	0	0
Unexpected pathogen	-1	-1
Additional unexpected pathogen	-1	-1

Pathogens are classified as ‘core’ if they can be readily identified in-house without the requirement for specialist methodology or expertise. ‘Advanced’ pathogens are those that would normally referred to a reference or expert laboratory for confirmation and/or specialist testing

Evaluation related to species and genus identification in bacteriology samples

● Basic scores

- **Hit score** (Schilling et al., 2006 [3])

- Value ranging between 0 (worst case) and 1 (best case) obtained by weighting the correct result according to the taxonomical identification of genus and species (qualitative) as well as the number of taxonomic specimens found (quantitative) in artificial samples.



Basic scores

• Hit score

A set of 16 identical artificial macrozoobenthos samples with 22 species in different numbers was prepared and sent to 16 laboratories.

Each sample included the following species:

Taxonomical group	Tested Species	Individuals per sample	Max. size of organisms (mm)	
Mollusca	<i>Arctica islandica</i>	1	3	
	<i>Corbula gibba</i> *	2	8	
	<i>Macoma balthica</i>*	51	17	
	<i>Mya sp.</i>	8	25	
	<i>Mysella bidentata</i> *	4	4	
	<i>Mytilus edulis</i> *	64	5	
	<i>Retusa obtusa</i>	1	6	
	<i>Tridonta borealis</i>	3	15	
	Polychaeta	<i>Eteone longa</i> *	1	40
		<i>Fabricia stellaris</i>	10	3
<i>Lagis koreni</i> *		2	30	
<i>Nephtys hombergii</i>		1	70	
<i>Pholoe assimilis</i>		2	7	
<i>Polydora quadrilobata</i>		1	20	
<i>Pygospio elegans</i> *		3	15	
<i>Scoloplos armiger</i> *		4	50	
<i>Terebellides stroemi</i> *		16	40	
Crustacea		<i>Corophium crassicorne</i>	1	7
	<i>Diastylis rathkei</i> *	87	20	
	<i>Gastrosaccus spinifer</i>	1	25	
	<i>Microdeutopus gryllotalpa</i>	1	9	
	<i>Phoxocephalus holbolli</i> *	1	12	

22

Phylum	Mollusca
Class	Bivalvia
Order	Cardiida
Family	Tellinidae
Genus	<i>Macoma</i> Leach, 1819
Species	<i>Macoma balthica</i> (Linnaeus, 1758)

Qualitative analysis
Quantitative analysis

Basic scores

Hit score

Table 2 Overview of qualitative and quantitative as well as combined successful hits and the corresponding categories for the assessment

Successful hits		Hit rate	Categorisation of hits
Qualitative analysis		1	Genus und species correct
		1-1/6	Genus correct, <u>species not named</u> ("sp.")
		1-1/3	Genus correct, <u>species false</u>
	1-1/3-1/3	0.333	<u>Genus and species false</u> , next taxonomical level correct
		0	Next taxonomical level false or species not found
Quantitative analysis		1	Species number correct
	1-1/6	0.833	Species number false
Combined qualitative/quantitative analysis		1	Genus und species correct, species number correct
	1-1/6	0.833	Genus correct, <u>species not named</u> ("sp."), species number correct
	1-1/3	0.667	Genus correct, <u>species false</u> , species number correct
		0.500	Genus und species correct, species number false
		0.417	Genus correct, species not named ("sp."), species number false
	1-1/3-1/3	0.333	Genus correct, <u>species false</u> , <u>species number false</u>
		0.333	Genus and species false, next taxonomical level correct, species number correct
	0.167	Genus and species false, next taxonomical level correct, species number false	
	0	Next taxonomical level false or species not found	

● Basic scores


- **Hit score**

- The arithmetic mean of the **hit scores of a laboratory** regarding all species **reflects its competence**
- The arithmetic mean of the **hit scores for a species** regarding all laboratories indicates the **difficulty in identifying individual species**

Elaborate scores

- **Results categorisation** (Clark and Wilson, 2005 [4])

- Presence/absence of drugs in 18 oral fluid samples



Sample	Amphetamines	Barbiturates	Cannabinoids	Methadone	Opiates
1	+	-	+		+	-
2	+	-	-		+	-
3	-	-	+		+	-
4	-	-	-		+	
5	-	+	+		-	+
.....
18	-	+	-		-	+

Elaborate scores

Results categorisation

- For each drug, the participants results are categorized as:
 - TP: True positive
 - TN: True negative
 - FP: False positive
 - FN: False negative
- Diagnostic sensitivity (DSe) and specificity (DSp) are calculated as function of TP, TN, FP and FN
- This categorization allows:
 - the identification of analytes with poor DSe or DSp
 - the identification of laboratories with poor performance due to the low sensitivity of the applied test

		Assigned value	
		+	-
Laboratory result	+	TP	FP
	-	FN	TN
		TP+FN	FP+TN

$$DSe = 100 \cdot \frac{TP}{TP + FN}$$

$$DSp = 100 \cdot \frac{TN}{TN + FP}$$

Elaborate scores

Results categorisation (Chabirand et al., 2014 [5])

- Presence/absence of plant pathogens in samples of plant matrices
- The participants results are categorized as:
 - PA: Positive agreement
 - NA: Negative agreement
 - PD: Positive deviation
 - ND: Negative deviation



Table 1 Definitions of the parameters of positive agreement (PA), negative agreement (NA), positive deviation (PD) and negative deviation (ND) (definitions adapted from ISO 16140)

Laboratory result	Assigned value	
	Positive	Negative
Positive	PA = positive agreement	PD = positive deviation
Negative	ND = negative deviation	NA = negative agreement
Indeterminate	ND = negative deviation	PD = positive deviation

Elaborate scores

Results categorisation

- The participant performance is evaluated as **trueness** and **precision**

Trueness: evaluated through the capacity to obtain positive results from positive samples (**sensitivity, Se**) and negative results from negative samples (**specificity, Sp**).

$$Se = 100 \cdot \frac{PA}{N^+}$$

$$Sp = 100 \cdot \frac{NA}{N^-}$$

$$AC = 100 \cdot \frac{PA+NA}{N} \quad N = N^+ + N^-$$

Laboratory result	Assigned value	
	Positive	Negative
Positive	PA = positive agreement	PD = positive deviation
Negative	ND = negative deviation	NA = negative agreement
Indeterminate	ND = negative deviation	PD = positive deviation
	N ⁺	N ⁻

Elaborate scores

- Results categorization

- The participant performance is evaluated as **trueness** and **precision**

Precision: evaluated through the capacity to obtain the same qualitative results from **identical samples analysed under condition of repeatability.**

$$DA = 100 \cdot \frac{PA + NA}{N}$$

Laboratory result	Assigned value	
	Positive	Negative
Positive	PA = positive agreement	PD = positive deviation
Negative	ND = negative deviation	NA = negative agreement
Indeterminate	ND = negative deviation	PD = positive deviation

Accordance (DA): Closeness of agreement between independent test results obtained under conditions of repeatability, i.e. conditions under which independent test results are obtained by the same method, **on identical test samples** in the same laboratory, by the same operator, using the same equipment, within a short period of time

● Elaborate scores

- **α -score** (Beavis et al., 2019 [6])

- α -score for qualitative testing that mimics the z-score commonly used for quantitative results
- Calculated as:

$$\alpha - score = I_c \cdot \frac{x - x_{pt}}{\sigma_{pt}}$$

where

x = participant result

x_{pt} = assigned value, obtained as estimated (consensus) outcome \hat{p} (**proportion** of satisfactory results)

σ_{pt} = fixed standard deviation for performance assessment

$$I_c \begin{cases} 1 & \text{if the consensus is "detected"} \\ -1 & \text{if the consensus is "not detected"} \end{cases}$$

Elaborate scores

- α -score**

Expected results for HIP 2: detected

\hat{p} (proportion of satisfactory results)=0.9643

\hat{q} (proportion of unsatisfactory results)=0.03571

$x_{pt} = \hat{p} = 0.9643$ $x = \hat{p}$ if the result is correct
 $x = \hat{q}$ if the result is incorrect

$$a - \text{score for lab01} = 1 \cdot \frac{0.9643 - 0.9643}{0.0524} = 0$$

$$a - \text{score for lab25} = 1 \cdot \frac{0.03571 - 0.9643}{0.0524} = -17.6$$

9 different species of highly infectious pathogens
 + correct results; - incorrect results

Lab	Species									ROD _{lab}	
	HIP 1	HIP 2	HIP 3	HIP 4	HIP 5	HIP 6	HIP 7	HIP 8	HIP 9		
01	+	+	+	+	-	+	-	-	-	56 %	5/9
02	+	+	+	+	+	+	+	+	+	100 %	9/9
03	+	+	+	+	-	+	-	-	+	78 %	
04	+	+	+	+	+	+	+	+	+	100 %	
05	+	+	+	+	-	+	+	+	+	89 %	
06	+	+	+	+	+	+	+	+	+	100 %	
07	+	+	+	-	+	+	-	-	+	67 %	
08	+	+	+	+	-	+	+	+	+	89 %	
09	+	+	+	+	+	+	+	+	+	100 %	
10	+	+	+	+	+	+	+	+	+	100 %	
11	+	+	+	+	+	+	+	+	+	100 %	
12	+	+	+	+	-	+	+	+	+	89 %	
13	+	+	+	+	-	+	+	+	+	89 %	
14	+	+	+	+	+	+	+	+	+	100 %	
15	+	+	+	+	-	+	+	+	+	89 %	
16	+	+	+	+	-	+	+	+	+	89 %	
17	+	+	+	+	+	+	+	+	+	100 %	
18	+	+	+	+	+	+	+	+	+	100 %	
19	+	+	+	+	+	+	+	+	+	100 %	
20	+	+	-	+	+	+	+	+	+	89 %	
21	+	+	+	+	+	+	+	+	+	100 %	
22	+	+	+	+	-	+	+	-	-	67 %	
23	+	+	+	+	+	+	+	+	-	89 %	
24	+	+	+	+	+	+	+	+	+	100 %	
25	+	-	+	-	+	+	+	+	+	78 %	
26	+	+	+	+	+	+	+	+	+	100 %	
27	+	+	+	+	+	+	+	+	+	100 %	
28	+	+	+	-	+	+	-	+	-	67 %	
ROD _{species}	100 %	96 %	96 %	89 %	68 %	100 %	89 %	86 %	86 %		

Elaborate scores

- α -score**

Expected results for HIP 2: detected


\hat{p} (proportion of satisfactory results)=0.9643

\hat{q} (proportion of unsatisfactory results)=0.03571

$x_{pt} = \hat{p} = 0.9643$ $x = \hat{p}$ if the result is correct
 $x = \hat{q}$ if the result is incorrect

$$a - \text{score for lab01} = 1 \cdot \frac{0.9643 - 0.9643}{0.0524} = 0$$

$$a - \text{score for lab25} = 1 \cdot \frac{0.03571 - 0.9643}{0.0524} = -17.6$$



Lab	HIP 1	HIP 2	HIP 3	HIP 4	HIP 5	HIP 6	HIP 7	HIP 8	HIP 9	SA2(a)	SA2(b)
Species											
01	0.0	0.0	0.0	0.0	-6.9	0.0	-14.9	-13.7	-13.7	71.7	74.8
02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
03	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	-13.7	0.0	26.1	23.6
04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
05	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	0.0	0.0	5.3	0
06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
07	0.0	0.0	0.0	-14.9	0.0	0.0	-14.9	-13.7	0.0	70.2	78.9
08	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	0.0	0.0	5.3	0
09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
12	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	0.0	0.0	5.3	0
13	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	0.0	0.0	5.3	0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
15	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	0.0	0.0	5.3	0
16	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	0.0	0.0	5.3	0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
20	0.0	0.0	-17.6	0.0	0.0	0.0	0.0	0.0	0.0	34.4	38.5
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
22	0.0	0.0	0.0	0.0	-6.9	0.0	0.0	-13.7	-13.7	47.0	47.2
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-13.7	20.9	23.6
24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
25	0.0	-17.6	0.0	-14.9	0.0	0.0	0.0	0.0	0.0	59.1	66.2
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
28	0.0	0.0	0.0	-14.9	0.0	0.0	-14.9	0.0	-13.7	70.2	78.9

Elaborate scores

- **α -score**

Interpretation:

- All results receiving non-zero scores can be considered unsatisfactory
- The usual confidence interval (± 2 and ± 3) associated to z-score can not be used
- A measure of “**how unsatisfactory**” is required in order to interpret α -scores. \bar{q} , a robust estimate for the mean q (proportion of unsatisfactory results) could provide it.
- Any value of q less than 3 SDs away from the mean are considered unsatisfactory

$$q < \bar{q} + 3\sigma_{pt}$$

$$a = I_c \cdot \frac{x - x_{pt}}{\sigma_{pt}}$$

- In the paper $q = 0.1986$ and $\alpha = 11.5$

Performance
evaluation

$$a = 0$$

Satisfactory

$$|a| < 11.50$$

Questionable

$$|a| \geq 11.50$$

Unsatisfactory

Advanced scores

- **K of Cohen-Fleiss** (Mancin et al., 2015 [7])

- K index can be used for binary (presence/absence) answers as well as for categorical/nominal answers (*Salmonella* serotyping, degree of disease...)
- **Cohen K**: An index of the agreement between the results of each participant and the assigned results → **performance participant evaluation**



	Sample 1	Sample 2	Sample 3	Sample 4	Sample 19	Sample 20	
Assigned results	+	+	-	-	+	-	➔ K lab 1
Lab 1 results	+	-	-	-	+	+	
.....							
Assigned results	+	+	-	-	+	-	➔ K lab 2
Lab 2 results	+	-	-	-	+	+	
.....							
Assigned results	+	+	-	-	+	-	➔ K lab 50
Lab 50 results	+	-	-	-	+	+	

Advanced scores

- K of Cohen-Fleiss

- Fleiss K: An index of the agreement among the results of the participants

➔ overall PT evaluation

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 19	Sample 20
Lab 1 results	+	-	-	-	+	+
Lab 2 results	+	+	+	-	+	-
.....	-	+	-	-	+	-
.....	+	+	+	-	+	-
Lab 50 results	+	-	-	-	-	-

Overall K

Advanced scores

• K of Cohen-Fleiss

- K relates the observed agreement (OA) with the expected agreement (EA) taking into account the agreement due to chance (1-EA)

$$K = \frac{(OA - EA)}{(1 - EA)}$$

- A different weight can be attributed to incorrect answers: weighted K

- A value of significance of K is available $z = \frac{\hat{K}}{s.e._0(\hat{K})} \approx N(0,1)$

- Landis and Koch provided a scale for the agreement interpretation:
< 0 “Poor”; 0.01-0.20 “Slight”; 0.21-0.40 “Fair”; 0.41-0.60 “Moderate”;
0.61-0.80 “Substantial”; 0.81-1.00 “Almost perfect”.

Advanced scores

- **Maximum likelihood estimation** (Schilling et al. 2006 [3], Uhlig et al. 2015 [8], Bashkansky et al. 2016 [9])
 - For presence/absence: more than one sample
 - All authors define the probability of success (p) as a function of **competence level (LCL)** and **level of difficulty (LDT)**
 - Logistic model is predominant in the case of binary answers

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = LCL - LDT$$

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 + \exp(LCL_i - LDT_j)}$$

- The probability of success for the laboratory i and sample j increases with the increase of the ability of the i laboratory (LCL_i) and with the decrease of difficulty of the sample j (LDT_j)

Advanced scores

- **Maximum likelihood estimation (MLE)** is a method to estimate the parameters of a statistical model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.

- Model
$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

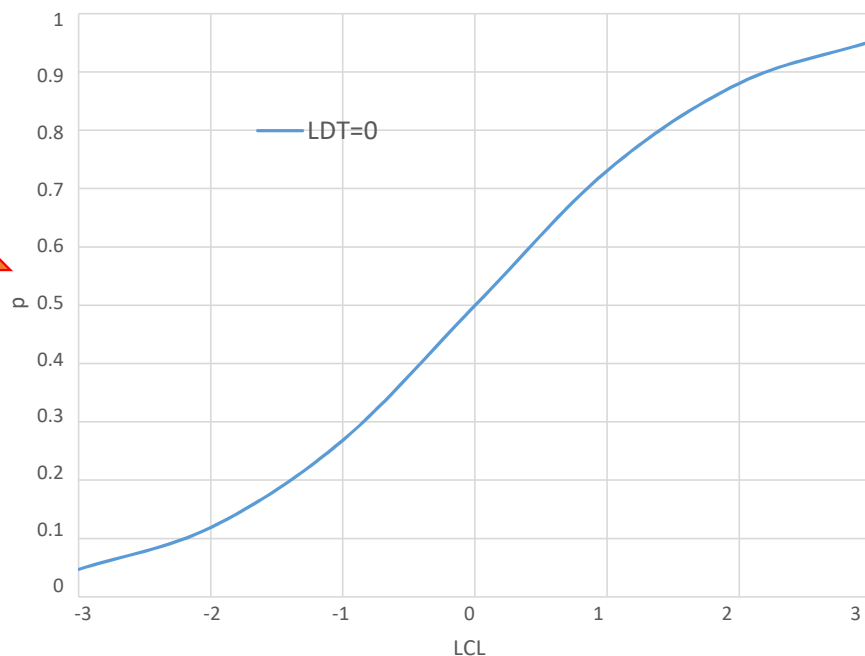
- Likelihood
$$l(\beta) = \sum_{i=1}^n y_i \beta x_i - \log(1 + e^{\beta x_i})$$

- Parameters
$$\widehat{\beta}_0 + \widehat{\beta}_1$$

$$\widehat{p} = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}$$

Advanced scores

- Maximum likelihood estimation (Schilling et al. 2006 [3], Uhlig et al. 2015 [8])



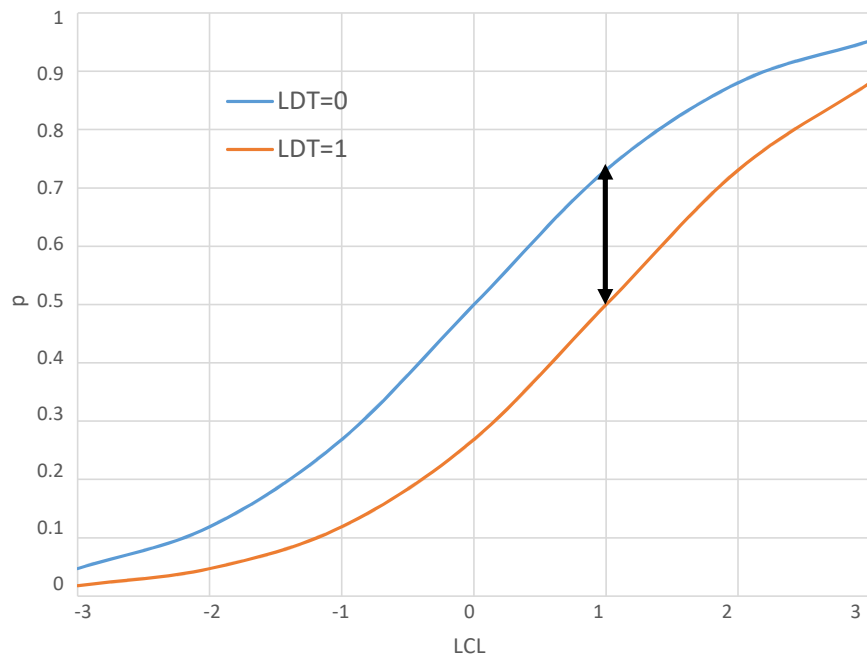
$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = LCL - \cancel{LDT} = LCL$$

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 + \exp(LCL_i - LDT_j)} = \frac{\exp(LCL_i)}{1 + \exp(LCL_i)}$$

In case of **no difficulty**, the probability of success is function only of the competence. The probability of success increases if the level of competence increases

Advanced scores

Maximum likelihood estimation



Difficulty increases LDT=1

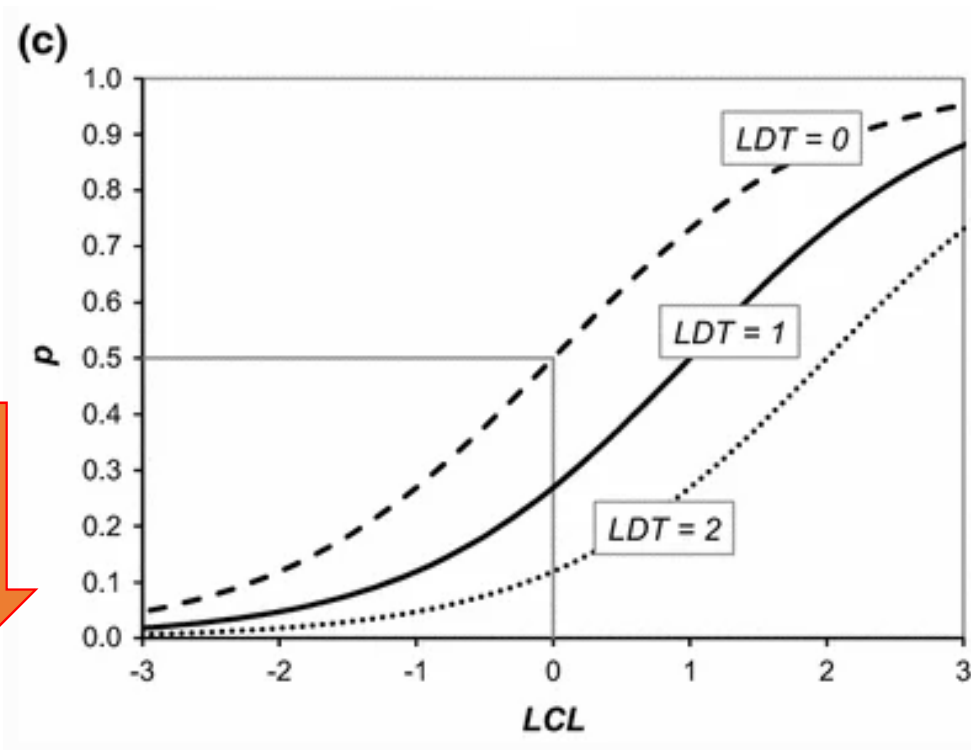


Probability of success decreases
(negative sign of LDT)

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 - \exp(LCL_i - LDT_j)}$$

Advanced scores

- Maximum likelihood estimation



$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = LCL - LDT = -LDT$$

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 + \exp(LCL_i - LDT_j)} = \frac{\exp(-LDT_j)}{1 + \exp(-LDT_j)}$$

Focusing on the mean level of competence ($LCL=0$), the probability of success decreases if the difficulty increases

Advanced scores

- Maximum likelihood method (Schilling et al. 2006 [3])

Lab i	$\ln \frac{h_i}{1-h_i}$	Prob. of success h_i	T_i
15	0.231355	0.557582	-1.95
4	0.57913	0.640867	-1.7
8	0.57913	0.640867	-1.7
16	0.57913	0.640867	-1.7
11	2.373649	0.914796	-0.41
3	2.763157	0.940652	-0.13
10	2.763157	0.940652	-0.13
13	2.763157	0.940652	-0.13
2	3.180487	0.960093	0.17
14	3.180487	0.960093	0.17
9	3.63955	0.974408	0.5
12	3.63955	0.974408	0.5
1	4.168168	0.984755	0.88
6	4.821985	0.992014	1.35
7	5.670556	0.996566	1.96
5	6.213085	0.998001	2.35

$$p_i = h_i = \frac{\exp(LCL_i - LDT)}{1 - \exp(LCL_i - LDT)}$$

average success value of a lab i considering a common level of difficulty of all species

Performance evaluation

$$T_i = \frac{\left(\ln \frac{h_i}{1-h_i} - \ln \frac{h_{median}}{1-h_{median}} \right)}{s}$$

$$h_{median} = \text{median}(h_i) = 0.9503$$

$$s = eMAD = 1.4826 MAD \left(\ln \frac{h_i}{1-h_i} \right) = 1.39112$$

MAD=median absolute deviation from the median

$T_i < -1.65$ laboratory is significantly poorer than the median laboratory

$T_i > 1.65$ laboratory is significantly better than the median laboratory

1.65, value corresponding to 95th percentile of normal distribution

Advanced scores

- Maximum likelihood method (Uhlig et al. 2015 [8])

Lab	Species									$p_{average}$
	HIP 1	HIP 2	HIP 3	HIP 4	HIP 5	HIP 6	HIP 7	HIP 8	HIP 9	
01	+	+	+	+	-	+	-	-	-	0.784
02	+	+	+	+	+	+	+	+	+	0.997
03	+	+	+	+	-	+	+	-	+	0.928
04	+	+	+	+	+	+	+	+	+	0.997
05	+	+	+	+	-	+	+	+	+	0.969
06	+	+	+	+	+	+	+	+	+	0.997
07	+	+	+	-	+	+	-	-	+	0.868
08	+	+	+	+	-	+	+	+	+	0.969
09	+	+	+	+	+	+	+	+	+	0.997
10	+	+	+	+	+	+	+	+	+	0.997
11	+	+	+	+	+	+	+	+	+	0.997
12	+	+	+	+	-	+	+	+	+	0.969
13	+	+	+	+	-	+	+	+	+	0.969
14	+	+	+	+	+	+	+	+	+	0.997
15	+	+	+	+	-	+	+	+	+	0.969
16	+	+	+	+	-	+	+	+	+	0.969
17	+	+	+	+	+	+	+	+	+	0.997
18	+	+	+	+	+	+	+	+	+	0.997
19	+	+	+	+	+	+	+	+	+	0.997
20	+	+	-	+	+	+	+	+	+	0.969
21	+	+	+	+	+	+	+	+	+	0.997
22	+	+	+	+	-	+	+	-	-	0.868
23	+	+	+	+	+	+	+	+	-	0.969
24	+	+	+	+	+	+	+	+	+	0.997
25	+	-	+	-	+	+	+	+	+	0.928
26	+	+	+	+	+	+	+	+	+	0.997
27	+	+	+	+	+	+	+	+	+	0.997
28	+	+	+	-	+	+	-	+	-	0.868
ROD_{species}	100 %	96 %	96 %	89 %	68 %	100 %	89 %	86 %	86 %	

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = LCL - LDT$$

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 + \exp(LCL_i - LDT_j)}$$

probability of success for lab i and sample j

$$p_{average} = \frac{\exp(LCL - \overline{LDT})}{1 + \exp(LCL - \overline{LDT})}$$

probability of success for a lab with competence level LCL and for a task with average level of difficulty \overline{LDT} .
 \overline{LDT} denotes the mean across all LDT_j parameter estimates.

Advanced scores

- Maximum likelihood method

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 - \exp(LCL_i - LDT_j)}$$

Lab	Species									Lab-specific L_j -score	$P_{average}$
	HIP 1	HIP 2	HIP 3	HIP 4	HIP 5	HIP 6	HIP 7	HIP 8	HIP 9		
01	0.0	0.0	0.0	0.1	-1.1	0.0	-1.9	-1.7	-1.7	-2.2	0.784
02	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
03	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	-1.7	0.1	-1.3	0.928
04	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
05	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
06	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
07	0.0	0.0	0.0	-1.9	0.3	0.0	-1.9	-1.7	0.1	-1.8	0.868
08	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
09	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
10	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
11	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
12	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
13	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
14	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
15	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
16	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
17	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
18	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
19	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
20	0.0	0.0	-2.3	0.1	0.3	0.0	0.1	0.1	0.1	-0.6	0.969
21	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
22	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	-1.7	-1.7	-1.8	0.868
23	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	-1.7	-0.6	0.969
24	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
25	0.0	-2.3	0.0	-1.9	0.3	0.0	0.1	0.1	0.1	-1.3	0.928
26	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
27	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
28	0.0	0.0	0.0	-1.9	0.3	0.0	-1.9	0.1	-1.7	-1.8	0.868

Performance evaluation

$$L_{ij} = \begin{cases} \Phi^{-1} \left(1 - \frac{0.5 \cdot \exp(-LDT_j)}{1 + \exp(-LDT_j)} \right) & \text{if laboratory } i \text{ successfully completed task } j \\ \Phi^{-1} \left(\frac{0.5}{1 + \exp(-LDT_j)} \right) & \text{if laboratory } i \text{ failed task } j \end{cases}$$

Φ denotes the cumulative function of the standard normal distribution

L_{ij} score for laboratory (i) and sample (j) takes into account both the level of difficulty of the task and the specific laboratory results.

A poor score ($L_{ij} < -2$) is only possible if the probability of obtaining an incorrect result is less than 5% (very easy task, low LDT)

A good score ($L_{ij} > 2$) is only possible if the probability of obtaining a correct result is less than 5% (very difficult task, high LDT)

Advanced scores

- Maximum likelihood method

$$p_{ij} = \frac{\exp(LCL_i - LDT_j)}{1 - \exp(LCL_i - LDT_j)}$$

Lab	Species									Lab-specific L_i score	$P_{average}$
	HIP 1	HIP 2	HIP 3	HIP 4	HIP 5	HIP 6	HIP 7	HIP 8	HIP 9		
01	0.0	0.0	0.0	0.1	-1.1	0.0	-1.9	-1.7	-1.7	-2.2	0.784
02	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
03	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	-1.7	0.1	-1.3	0.928
04	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
05	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
06	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
07	0.0	0.0	0.0	-1.9	0.3	0.0	-1.9	-1.7	0.1	-1.8	0.868
08	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
09	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
10	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
11	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
12	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
13	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
14	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
15	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
16	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	0.1	0.1	-0.6	0.969
17	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
18	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
19	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
20	0.0	0.0	-2.3	0.1	0.3	0.0	0.1	0.1	0.1	-0.6	0.969
21	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
22	0.0	0.0	0.0	0.1	-1.1	0.0	0.1	-1.7	-1.7	-1.8	0.868
23	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	-1.7	-0.6	0.969
24	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
25	0.0	-2.3	0.0	-1.9	0.3	0.0	0.1	0.1	0.1	-1.3	0.928
26	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
27	0.0	0.0	0.0	0.1	0.3	0.0	0.1	0.1	0.1	1.1	0.997
28	0.0	0.0	0.0	-1.9	0.3	0.0	-1.9	0.1	-1.7	-1.8	0.868

Performance evaluation

$$L_i = \frac{\exp(LCL_i - \overline{LCL})}{SE(LCL_i)}$$

where SE is the standard error and \overline{LCL} is the average level of competence. \overline{LCL} denotes the mean across all LCL_i parameter estimates

- $L_i < -2$ lower than the average competence
- $|L_i| \leq 2$ average competence
- $L_i > 2$ higher than the average competence

Advanced scores



- **ORDANOVA** Ordinal Analysis of Variance (Bashkansky et al. 2012, [10])
 - Ordinal results (according to magnitude, $K=1,2,3,4$) or binary results (as two level of scales, sex for human being, $K=1,2$)
 - h^2 index as function of cumulative frequency F_k

$$h^2 = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} F_k \cdot (1 - F_k)$$

to measure the degree of data variation in terms of

Total variation in a PT $\hat{h}_{(T)}^2$	Within-laboratory Variation $\hat{h}_{m(W)}^2$	Between—laboratories variation $\hat{S}_{k(B)}^2$
$\hat{h}_{(T)}^2 = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \hat{F}_k \cdot (1 - \hat{F}_k)$	$\hat{h}_{m(W)}^2 = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \hat{F}_{km} \cdot (1 - \hat{F}_{km})$	$\hat{S}_{k(B)}^2 = \frac{1}{m} \sum_{m=1}^M (\hat{F}_{km} - \hat{F}_k.)^2$

Advanced scores

ORDANOVA

- N=100 similar items belonging to the same category 2, were analysed from 3 laboratories: A, B, C
- Aim: to evaluate differences in the results of laboratories

Table 1 Test results (numbers n_k of results of category k and corresponding frequencies \hat{p}_k and \hat{F}_k) obtained by different laboratories in case 1

Laboratory	A			B			C		
	Category k	n_k	\hat{p}_k	\hat{F}_k	n_k	\hat{p}_k	\hat{F}_k	n_k	\hat{p}_k
1	4	0.04	0.04	2	0.02	0.02	3	0.03	0.03
2	83	0.83	0.87	90	0.90	0.92	85	0.85	0.88
3	10	0.10	0.97	6	0.06	0.98	7	0.07	0.95
4	3	0.03	1.00	2	0.02	1.00	5	0.05	1.00

n_k =number of results belonging to the k-th category
 n =total number of results
 p_k =proportion of results in the k-th category
 F_k = cumulative frequency

$$\hat{p}_k = \frac{n_k}{n} \quad \sum_{k=1}^K \hat{p}_k = 1$$

$$\hat{F}_k = \sum_{i=1}^k \hat{p}_i = 1 \quad \hat{F}_K = 1$$

Advanced scores

- ORDANOVA

Laboratory/ category	A	B	C	T		Between
1	0.04	0.02	0.03	0.03	$\hat{S}_{1(B)}^2$	6.67e-5
2	0.87	0.92	0.88	0.89	$\hat{S}_{2(B)}^2$	4.67e-4
3	0.97	0.98	0.95	0.97	$\hat{S}_{3(B)}^2$	1.56e-4
4	1.00	1.00	1.00	1.00	$\hat{S}_{4(B)}^2$	0
	$\hat{h}_{A(W)}^2$	$\hat{h}_{B(W)}^2$	$\hat{h}_{C(W)}^2$			
Within	0.2408	0.1504	0,2429			

\hat{F}_{km}

$\hat{S}_{(B)}^2 = 5,74e - 5$

$\hat{h}_{(W)}^2 = 0,21138$

$\hat{h}_{(T)}^2 = \hat{h}_{(W)}^2 + \hat{S}_{(B)}^2 = 0,2123$

● Advanced scores

• ORDANOVA

- Results from different laboratories are significantly different if:

$$\frac{\text{between lab variation}}{\text{total variation}} = \frac{\hat{S}_{(B)}^2}{\hat{h}_{(T)}^2} > \frac{df_B}{df_T} \quad df_B = M - 1; df_T = N - 1$$

- Defined the indicator I as

$$I = \frac{\hat{S}_{(B)}^2 / df_B}{\hat{h}_{(T)}^2 / df_T}$$

- The greater I exceeds unity, the greater is the difference in the results of the laboratories and *vice versa*

37

Advanced scores

• CATANOVA Two-way categorical analysis (Gadrich et al., 2020 [11])

- Generalization of ORDANOVA
 - Nominal data with more than two categories
 - Two factor variables (laboratories **AND** technician experience)
- Example: 12 images of welds with **5 categories of imperfection** as test items for examination. 3 laboratories and 2 technicians were involved

Table 1 Classification of features by macroscopic examination

Category/class of the weld imperfection	Feature—designation of the imperfection	Reference number of the feature
1	Cracks	100
2	Cavities	200
3	Inclusions	300
4	Lack of fusion/penetration	400
5	Geometrical shape errors	500

Advanced scores

- CATANOVA Two-way categorical analysis

Table of summary results for class of imperfections

Class	Laboratory						Total
	L1		L2		L3		
	A	B	A	B	A	B	
1	1	4	1	4	0	1	11
2	2	3	3	2	2	2	14
3	2	2	1	2	1	1	9
4	6	4	6	2	5	6	29
5	3	1	3	4	6	4	21
Total	14	14	14	14	14	14	84

Total variation in a PT

$$\hat{V}_{(T)}$$

Intra-laboratory

Variation \hat{V}_W

Inter-laboratories

Variation \hat{C}_B

$$\hat{V}_T = \frac{K}{K-1} \left(1 - \sum_{k=1}^K \hat{p}_{..k}^2 \right)$$

$$\hat{V}_W = \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} \cdot \frac{K}{K-1} \left(1 - \sum_{k=1}^K \hat{p}_{ijk}^2 \right)$$

$$\hat{C}_B = \frac{K}{K-1} \sum_{k=1}^K \sum_{i=1}^I \sum_{j=1}^J \pi_{ij} \cdot (\hat{p}_{ijk} - \hat{p}_{.k})^2$$

Advanced scores

- **CATANOVA Two-way categorical analysis**

- R^2 is the joint effect of the factors on the results

$$R^2 = \frac{\text{inter - lab variation}}{\text{total variation}} = \frac{\hat{C}_B}{\hat{V}_T}, \quad 0 \leq R^2 \leq 1$$

- I indicator defined as

$$\hat{I} = (IJ - 1)(K - 1)\widehat{SP}_B = (IJ - 1)(K - 1) \frac{\hat{C}_B/df_B}{\hat{V}_T/df_T}$$

where SP_B is the index of segregation power or index of dissimilarity.

- I allows to accept/reject the significance of the factors on the response variable

Advanced scores

- **CATANOVA Two-way categorical analysis**

- R^2 and I can be calculated to evaluate:

- The influence of both factors, laboratory (X1) and technician experience (X2)
- The influence of the individual effect of factors X1, X2 and their interaction $X1*X2$

on the variability of the obtained results

- The significance of I allows to conclude if:

- Both factors (laboratories and their technicians)
- Individual factor (technicians (X1) or laboratories (X2))
- Their interaction ($X1*X2$)

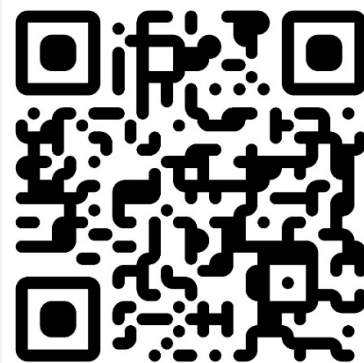
differ in the examination results for weld imperfections

● Summary

- I presented an overview of the basic/elaborate/advanced scores reported by Tiikkainen et al.
- Basic, elaborate or advanced scores can be used to provide a non quantitative PT performance evaluation

Are there other methods to evaluate the performance in non- quantitative PT?

Scan the QR code

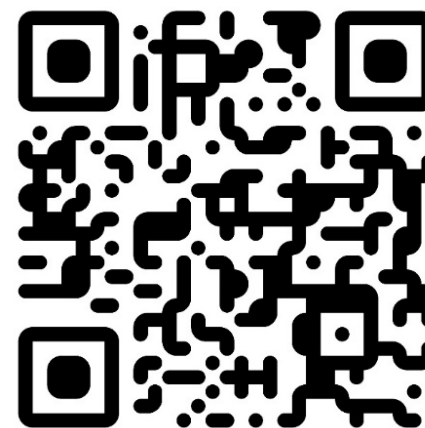


and contribute to the **2023 survey on non-quantitative PT!!**

Acknowledge

- Thanks to Ulla, Laura, Carolina, Markus, Marina, Piotr and Erika for this very significant ($p\text{-value} < 0.001$) work of review concerning the methods to evaluate the non-quantitative PT performances
- Thank you to the entire Eurachem PT group for these months of collaboration
- Thank you for your attention.

Scan the QR code



2023 survey on non-quantitative PT

References

- [1] <https://doi.org/10.1007/BF02014425>
- [2] <https://doi.org/10.1007/s00769-012-0895-1>
- [3] <https://doi.org/10.1007/s00769-006-0139-3>
- [4] <https://doi.org/10.1016/j.forsciint.2004.11.025>
- [5] <https://doi.org/10.1007/s00769-014-1034-y>
- [6] <https://doi.org/10.1007/s00769-019-01386-8>
- [7] <https://doi.org/10.1007/s00769-015-1129-0>
- [8] <https://doi.org/10.1007/s00769-015-1174-8>
- [9] <https://doi.org/10.1007/s00769-016-1208-x>
- [10] <https://doi.org/10.1007/s00769-011-0856-0>
- [11] <https://doi.org/10.1007/s42452-020-03907-4>