



## *Distribution of PT results and their comparability*

**Dr. Ilya Kuselman**

The National Physical Laboratory of Israel (INPL)  
Givat Ram, Jerusalem, e-mail: [ilya.kuselman@moital.gov.il](mailto:ilya.kuselman@moital.gov.il)  
<http://inpl.moital.gov.il>

Vice-Chair of the Co-operation on International Traceability  
in Analytical Chemistry (CITAC), <http://www.citac.cc>



## *Comparability and traceability concepts*

The concept of **comparability** (equivalence) of measurement results - “**tested once, accepted everywhere**” - is increasingly important since it allows to minimize technical barriers in trade and to cut down expense for international cooperation.

While results obtained under repeatability conditions can be compared directly, results obtained by different laboratories in different countries and at different times are comparable through their **relationship to the same reference** which is an internationally agreed and recognized measurement standard. This strategy is termed “**traceability**”.

Practical estimation of comparability of measurement/testing/analytical results is based on intercomparisons.

**Key comparisons** conducted on the best measurement capability level of NMIs are organized by CCQM at BIPM in the framework of the Convention of the Meter ([www.bipm.org](http://www.bipm.org)).

Intercomparisons organized on the routine measurement level of field laboratories are named “**proficiency testing**” (PT) since they are used mostly for assessment of a laboratory performance.

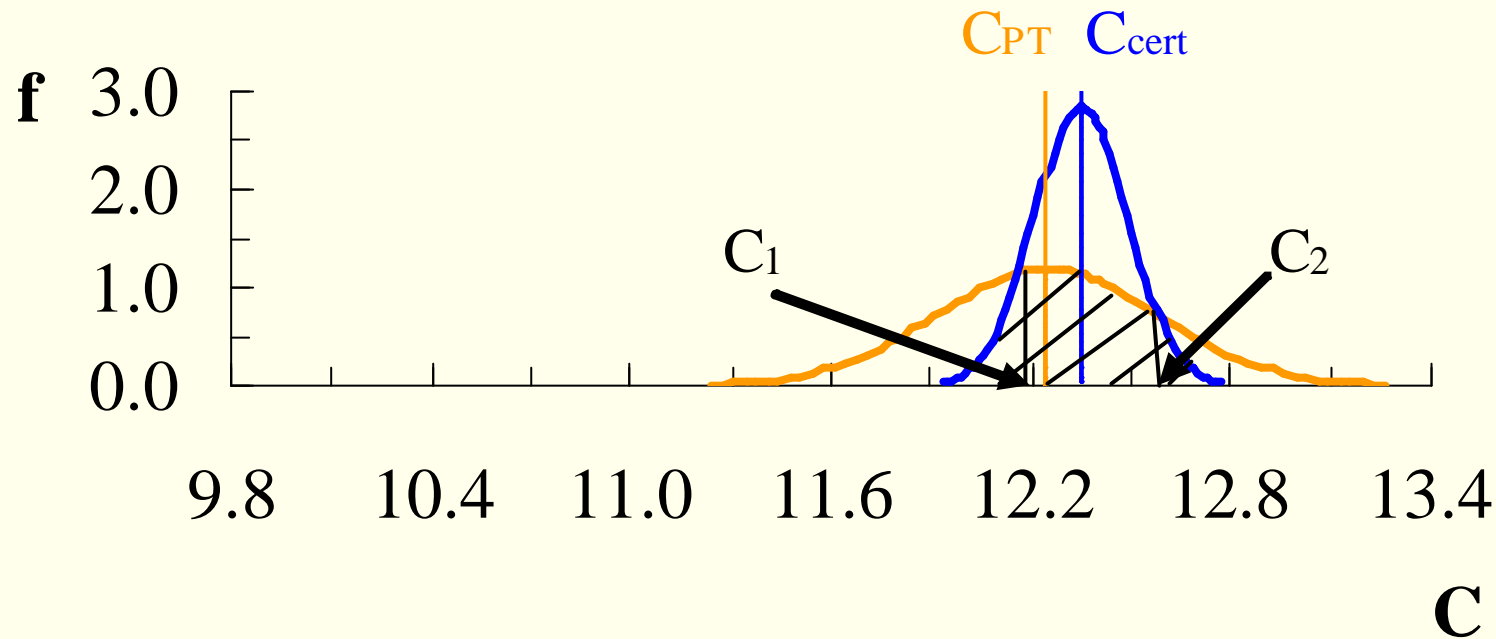
Two main steps are common for all PT schemes:

- 1) estimation of the **assigned value** of analyte concentration in the test items/RM and quantification of the value's **uncertainty** including components arising from the material homogeneity and stability, and
- 2) calculation of a laboratory performance statistics.

As test items, portions of a **CRM** - in the ideal case - are distributed among the laboratories participating in the PT. If CRM is expensive or not available, an **IHRM** or **a spike** with the traceable property value can be helpful.

One can imagine a situation when CRM, IHRM or a spike, with a traceable certified/assigned value of the analyte concentration  $C_{\text{cert}}$  and standard uncertainty  $\sigma_{\text{cert}}$ , is used for proficiency testing of (theoretically) **infinite population** of laboratories that produced results with the mean  $C_{\text{PT}}$  and standard deviation  $\sigma_{\text{PT}}$ .

For simplicity, the data of RM certification and PT results are considered as independent random events having both **normal distributions** with parameters  $C_{\text{cert}}$ ,  $\sigma_{\text{cert}}$  and with  $C_{\text{PT}}$ ,  $\sigma_{\text{PT}}$ , correspondingly.



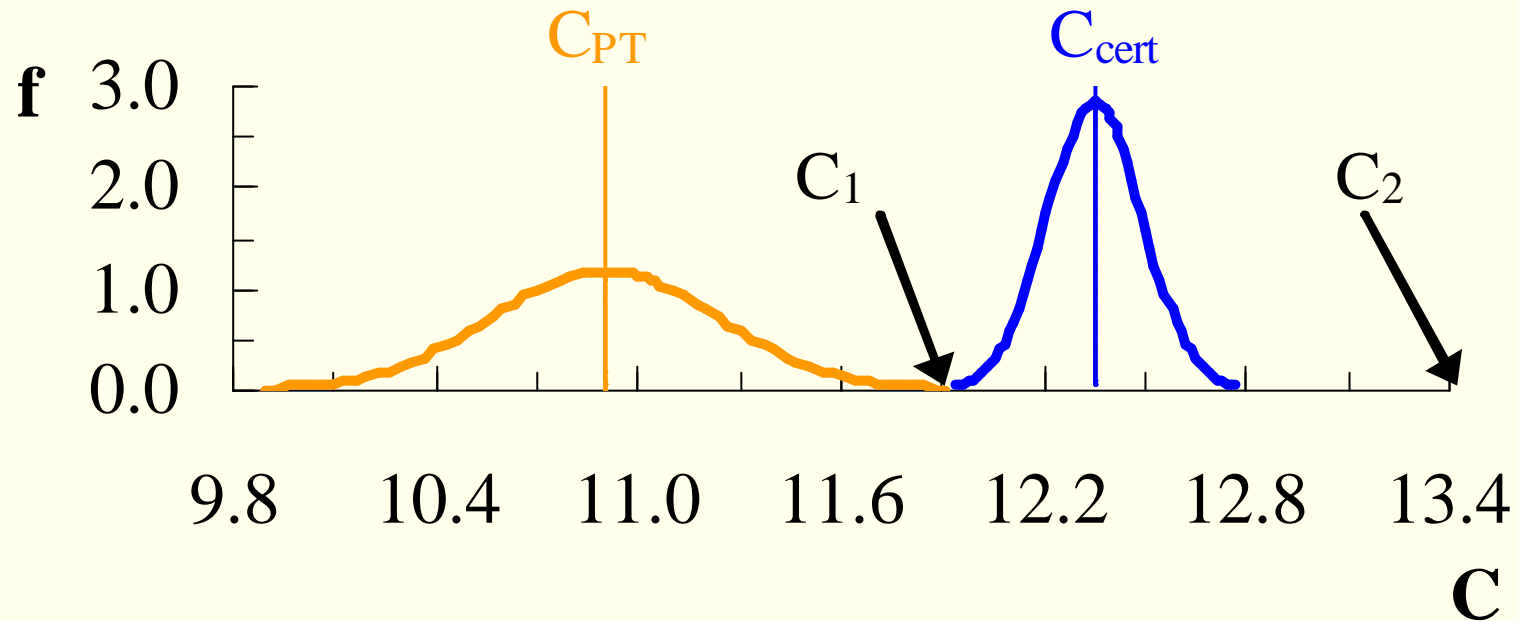
*Density functions (aluminum in fly ashes, % by weight, SRM 2690:  $C_{cert} = 12.35$  and  $s_{cert} = 0.14$ ) at  $C_{PT} = 12.25$  and  $s_{PT} = 0.34$ . The shaded area is the probability of the joint events.*

$$f_{PT} = \frac{1}{\mathbf{s}_{PT} \sqrt{2\mathbf{p}}} e^{-(C-C_{PT})^2/2\mathbf{s}_{PT}^2} = \frac{1}{\mathbf{s}_{cert} \sqrt{2\mathbf{p}}} e^{-(C-C_{cert})^2/2\mathbf{s}_{cert}^2} = f_{cert} \quad (1)$$

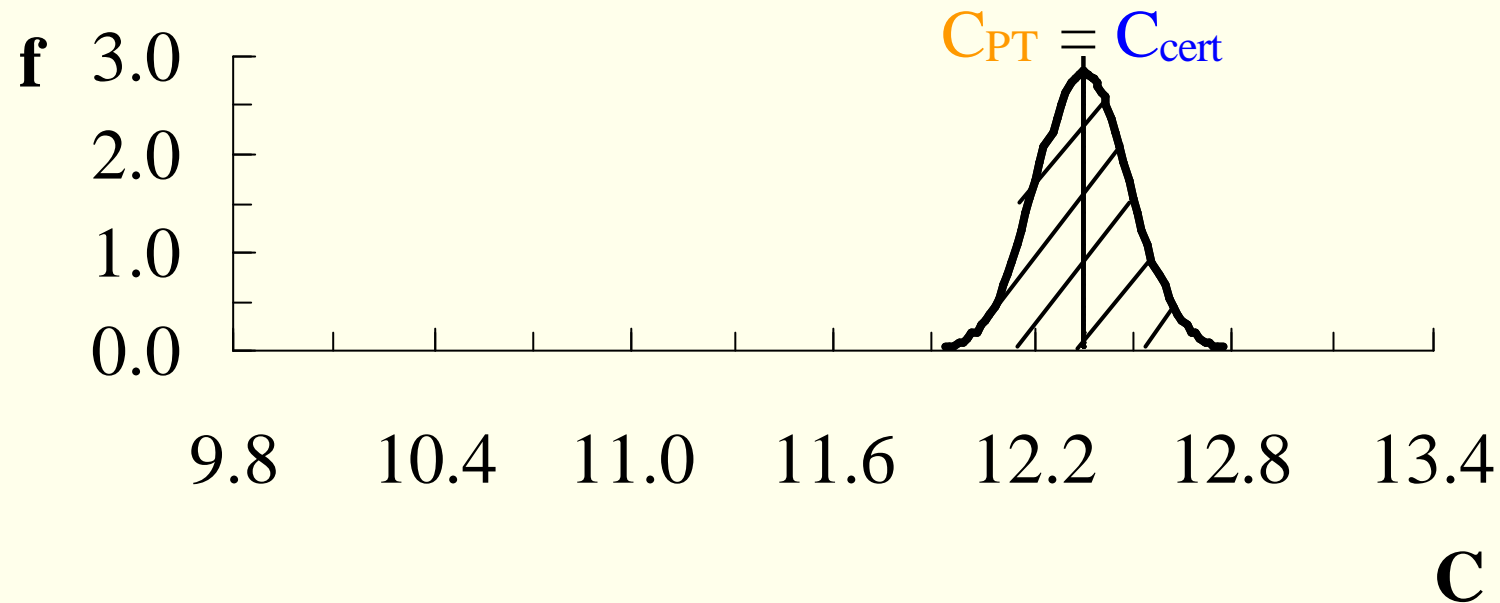
$$C_1, C_2 = \frac{(\mathbf{s}_{cert}^2 C_{PT} - \mathbf{s}_{PT}^2 C_{cert}) \pm \mathbf{s}_{cert} \mathbf{s}_{PT} \sqrt{\mathbf{r}}}{\mathbf{s}_{cert}^2 - \mathbf{s}_{PT}^2}, \quad (2)$$

$$\mathbf{r} = (C_{cert} - C_{PT})^2 + 2(\mathbf{s}_{PT}^2 - \mathbf{s}_{cert}^2) \ln \frac{\mathbf{s}_{PT}}{\mathbf{s}_{cert}}. \quad (3)$$

$$P = \int_{-\infty}^{C_1} f_{cert} dC + \int_{C_1}^{C_2} f_{PT} dC + \int_{C_2}^{+\infty} f_{cert} dC = 1 + \mathbf{f}\left(\frac{C_1 - C_{cert}}{\mathbf{s}_{cert}}\right) + \mathbf{f}\left(\frac{C_2 - C_{PT}}{\mathbf{s}_{PT}}\right) - \mathbf{f}\left(\frac{C_1 - C_{PT}}{\mathbf{s}_{PT}}\right) - \mathbf{f}\left(\frac{C_2 - C_{cert}}{\mathbf{s}_{cert}}\right) \quad (4)$$



*Density functions at  $C_{PT} = 10.9$  and  $S_{PT} = 0.34$*



*Any other values of distribution parameters lead to  $P < 1$ .  
Even at  $C_{PT} = C_{cert}$  and  $S_{PT} > S_{cert}$ , it is impossible to  
assess the comparability unambiguously.*

$H_0$  may consist of the assumption that the bias  $|C_{PT} - C_{cert}|$  exceeds  $\sigma_{cert}$  by a value which is **insignificant** in comparison with random interlaboratory errors of the analysis  $\sigma_{PT}$ . In this case, the hypothesis has the following form:

$$H_0: |C_{PT} - C_{cert}| = [(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2}.$$

At ratio  $\gamma = \sigma_{cert}/\sigma_{PT} = 0.4$   $H_0$  corresponds to probability  $P \geq 0.56$ , and at ratio  $\gamma = 0.7$  it corresponds to  $P \geq 0.77$ . When  $\sigma_{PT} = \sigma_{cert}$  and  $\gamma = 1.0$ , the probability values are  $P \geq 0.83$ .

The alternative hypothesis  $H_1$  assumes that the bias **exceeds**  $\sigma_{\text{cert}}$  significantly, e.g.

$$H_{11}: |C_{\text{PT}} - C_{\text{cert}}| = 2.0 \times [(0.3\sigma_{\text{PT}})^2 + \sigma_{\text{cert}}^2]^{1/2},$$

$$H_{12}: |C_{\text{PT}} - C_{\text{cert}}| = 2.1 \times [(0.3\sigma_{\text{PT}})^2 + \sigma_{\text{cert}}^2]^{1/2}, \text{ etc.}$$

In practice, laboratories participating in a PT form a **statistical sample** (from the population) of size  $N$ , i.e. only  $N$  laboratories are sending their results.

$H_0$  is not rejected when

$$| C_{PT/av} - C_{cert} | + t_{1-\alpha/2} S_{PT} / N^{1/2} = [(0.3\sigma_{PT})^2 + \sigma_{cert}^2]^{1/2},$$

where  $C_{PT/av}$  and  $S_{PT}$  are the sample estimates of  $C_{PT}$  and  $\sigma_{PT}$  calculated from the same  $N$  results; the left-hand side is the upper limit of the confidence interval for the bias

$| C_{PT} - C_{cert} |$ ;  $t_{1-\alpha/2}$  is the quantile of the Student's distribution for  $f = N-1$ ; the value  $1-\alpha/2$  is the probability of the bias  $| C_{PT} - C_{cert} |$  not exceeding its upper limit.



## The bias norms in $S_{PT}$ units at $a = 0.05$

$\gamma$	N						
	5	10	15	20	30	40	50
0.4	0.20	0.20	0.23	0.26	0.30	0.32	0.34
0.7	0.95	0.68	0.65	0.64	0.65	0.66	0.67
1.0	1.76	1.19	1.09	1.06	1.03	1.02	1.02

The value  $\gamma$  is set taking into account a  $\sigma_{PT}$  value equal to the standard analytical/measurement uncertainty or to the target standard deviation (calculated using the Horwitz curve or another database).

If SRM 2690 (NIST) is chosen for PT of Al determination in coal fly ashes,  $C_{\text{cert}} = 12.35\%$  by weight and  $\sigma_{\text{cer}} = 0.28/2 = 0.14\%$  ( $U = \pm 0.28\%$  is shown in the certificate).

The ASTM standard: “the means of the results of duplicate determinations carried out by different laboratories on riffled splits of the analysis sample should not differ by more than 2.0% for  $\text{Al}_2\text{O}_3$ ”, i.e. 1.06% for Al.

Since the range for two lab results is limited by the ASTM,  $\sigma_{\text{PT}} = 1.06/2.77 = 0.38\%$ , where 2.77 is the 95% percentile of the range distribution. The value  $\gamma = 0.14/0.38 = 0.4$ .



# PT results for Al in SRM 2690

Lab. No.	Result, %, w/w	Lab. No.	Result, %, w/w
1	12.76	16	12.60
2	12.19	17	12.81
3	12.68	18	12.39
4	12.21	19	11.96
5	12.96	20	11.91
6	12.27	21	11.86
7	11.96	22	12.32
8	12.03	23	12.53
9	11.88	24	12.84
10	11.97	25	12.67
11	12.23	26	12.86
12	12.48	27	12.75
13	12.69	28	12.66
14	12.21	29	11.99
15	11.98	30	12.61
$C_{PT/av}$	12.30	$C_{PT/av}$	12.38
$S_{PT}$	0.34	$S_{PT}$	0.35

Comparability of the results of 15 laboratories can be assessed as **satisfactory** since at  $\gamma = 0.14/0.38 = 0.4$

$$|C_{PT/av} - C_{cert}| = |12.30 - 12.35| = 0.05 < 0.23 \quad S_{PT} = 0.23 \times 0.34 = 0.08 \%$$

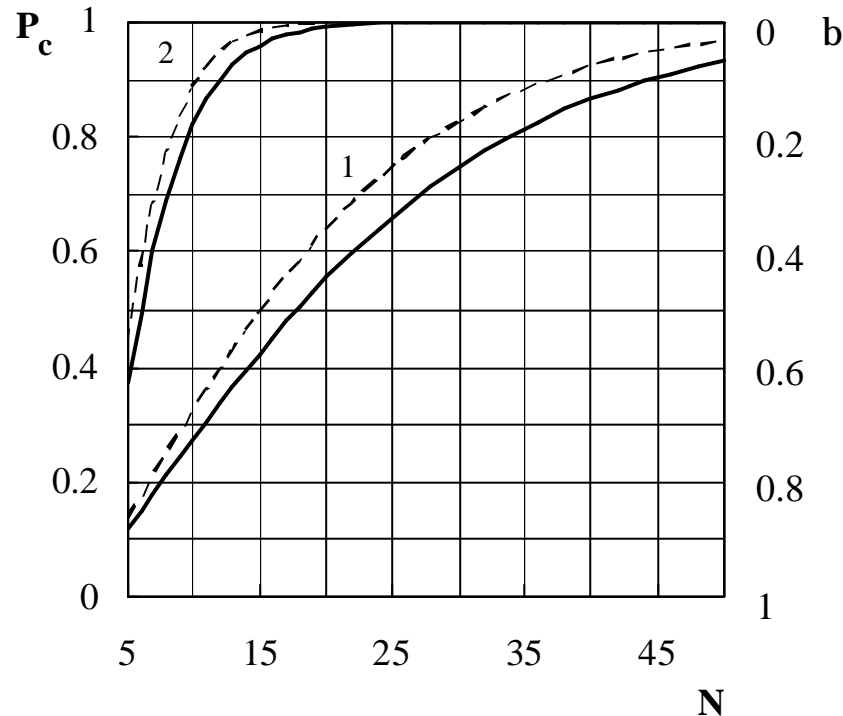
Comparability of the results of 30 laboratories is also assessed here as **satisfactory** since

$$|C_{PT/av} - C_{cert}| = |12.38 - 12.35| = 0.03 < 0.30 \quad S_{PT} = 0.30 \times 0.35 = 0.11\%$$

**Reliability** in such comparability assessment is determined by the probabilities

- of not rejecting the null hypothesis  $H_0$  when it is **true**, and
- of rejecting  $H_0$  when it is **false** (the alternative hypothesis  $H_1$  is true).

The criterion does not allow rejecting hypothesis  $H_0$  with probability  $1-\alpha/2$ , when it is true. Probability of an **error of type I** by this criterion (to reject the  $H_0$  hypothesis when it is true) is  $\alpha/2$ .



$\beta = 1 - P_c$  is the probability of an **error of type II** (not rejecting the  $H_0$  when it is false). Solid lines are for  $H_0$  testing against  $H_{11}$ , dotted lines – against  $H_{12}$ . Curves 1 and 2 are at  $\gamma = 0.4$  and at  $\gamma = 1.0$ .

I.Kuselman. *Accred. Qual. Assur.* (2006) 10: 466-470.



## *Chemical/metrological point of view*

Discussion of the comparability should be restricted by the definition of the analyte/matrix for which  $C_{\text{cert}}$  and  $\sigma_{\text{cert}}$  are quantified: **adequacy** of the RM used is very important.

When the RM prepared for PT is not certified and a **consensus value** (average or median of PT results) is used instead of  $C_{\text{cert}}$ , traceability of this value is questionable and comparability of the PT results cannot be assessed in the meaning “tested once, accepted everywhere”. In such cases, especially when number  $N$  of the participants is limited, a **local comparability**, i.e. among the participants only, is tested.

1. Comparability of PT results can be assessed based on a criterion of “yes-no”-type for testing  $H_0$  about insignificance of the bias of the results mean from the traceable certified RM value used for the PT.
2. A combination of chemical/metrological and statistical knowledge is necessary for careful formulation of  $H_0$ , since different hypotheses can lead to different decisions about comparability of the results obtained in the same PT scheme.

[I.Kuselman. Accred. Qual. Assur. \(2006\) 10: online](#)