

ISTA GMO task force



Bangkok, Ordinary meeting
2005

Rating ISTA Proficiency Tests
on GMO testing

GREGOIRE Sylvain LAFFONT Jean-Louis REMUND Kirk
HALDEMANN Christoph KAHLERT Bettina &Al.

Suggested decoding of rates = as for other proficiency tests

- A No problem has been detected in this test
- B There are small problems, but no specific look or action is suggested to the participant
- C Problems, ISTA indicates there might be things to consider by the laboratory to explain or correct things
- BMP Below Minimum of Performance, ISTA indicates that the results were poor and the laboratory need to find explanations and to improve/correct

Rating as for other proficiency tests

- Rating is a general feature of ISTA proficiency tests, provided by each technical committee through the test leader and ISTA secretariat, with the assistance of the proficiency test committee and statistics committee

Example: 5A rating and 1 BMP (Cotton)

*5*5points + 1*0point =25 points*

Overall rating is B

Example 2:

4 B =16 points

each proficiency test

and a run of 6 tests

One test rating	One test Score Value	Overall rating on 6 tests	Range on 6 tests
A	5 points	A	28 – 30 points
B	4 points	B	21 – 27 points
C	3 points	C	16 – 20 points
BMP	0 points	BMP	below 16 points

Information available

- Before to send the samples to the laboratories
 - Protocol (events, sample size,...)
 - Purity check (Conventional and GM seeds)
 - True value for each sample (spiking) *= better situation than GE,PU,OSD,Vigour,...*
- Result sheets returned from laboratories
 - Presence/absence
 - Quantification when presence
 - Other information (method, raw data points,...)
- Computational results by laboratory, and from all laboratories (median...)

Rating a laboratory in a given proficiency test

- We need to compute objectively, and transparently to the participants, criteria to rate laboratories for each proficiency test on:
 - Presence/absence
 - 3 systems are proposed to rate presence/absence
 - Quantification
 - 3 systems are proposed to rate quantification

Rating presence/absence

- *True value is known (purity check + more than 2 seeds are spiked)*
- 2 types of error can occur:
 - Laboratory report as negative a positive sample
 - Laboratory report as positive a negative sample



Presence/absence

3 suggested rating systems

Example with 12 samples in a Test

mistakes	% mistakes
1/12	8,3%
2/12	16,7%
3/12	25,0%
4/12	33,3%

Rate	Fixed Number of misclassified samples	Rate	Percentage of misclassified samples	Rate	Percentage of misclassified samples
A	0 errors	A	0% - 5%	A	0% - 6%
B	1 or 2 errors	B	>5% - 10%	B	>6% - 20%
C	3 errors	C	>10% - 20%	C	>20% - 30%
BMP	More than 3 errors	BMF	>20%	BMP	>30%

System1 System2 System3

3 mistakes $3/12=25\%$

Presence/absence

3 suggested rating systems

mistakes	% mistakes
1/12	8,3%
2/12	16,7%
3/12	25,0%
4/12	33,3%

Rate	Fixed Number of misclassified samples
A	0 errors
B	1 or 2 errors
C	3 errors
BMP	More than 3 errors

System1

Rate	Percentage of misclassified samples
A	0% - 5%
B	>5% - 10%
C	>10% - 20%
BMP	>20%

System2

Rate	Percentage of misclassified samples
A	0% - 6%
B	>6% - 20%
C	>20% - 30%
BMP	>30%

System3

Example with 12 samples in a Test : *system 2 is little more stringent*

Rate	Fixed Number of misclassified samples
A	0 errors
B	1 or 2 errors
C	3 errors
BMP	More than 3 errors

Rate	Percentage of misclassified samples
A	0 errors
B	1 error
C	2 errors
BMP	More than 2 errors

Rate	Percentage of misclassified samples
A	0 errors
B	1 or 2 errors
C	3 errors
BMP	More than 3 errors



Presence/absence

3 suggested rating systems

Mistakes	% mistakes
1/20	5.0%
2/20	10.0%
3/20	15.0%
4/20	20.0%
5/20	25.0%
6/20	30.0%

Rate	Fixed Number of misclassified samples
A	0 errors
B	1 or 2 errors
C	3 errors
BMP	More than 3 errors

System1

Rate	Percentage of misclassified samples
A	0% - 5%
B	>5% - 10%
C	>10% - 20%
BMP	>20%

System2

Rate	Percentage of misclassified samples
A	0% - 6%
B	>6% - 20%
C	>20% - 30%
BMP	>30%


System3

Example with 20 samples in a Proficiency Test: *System 3 is more lenient*

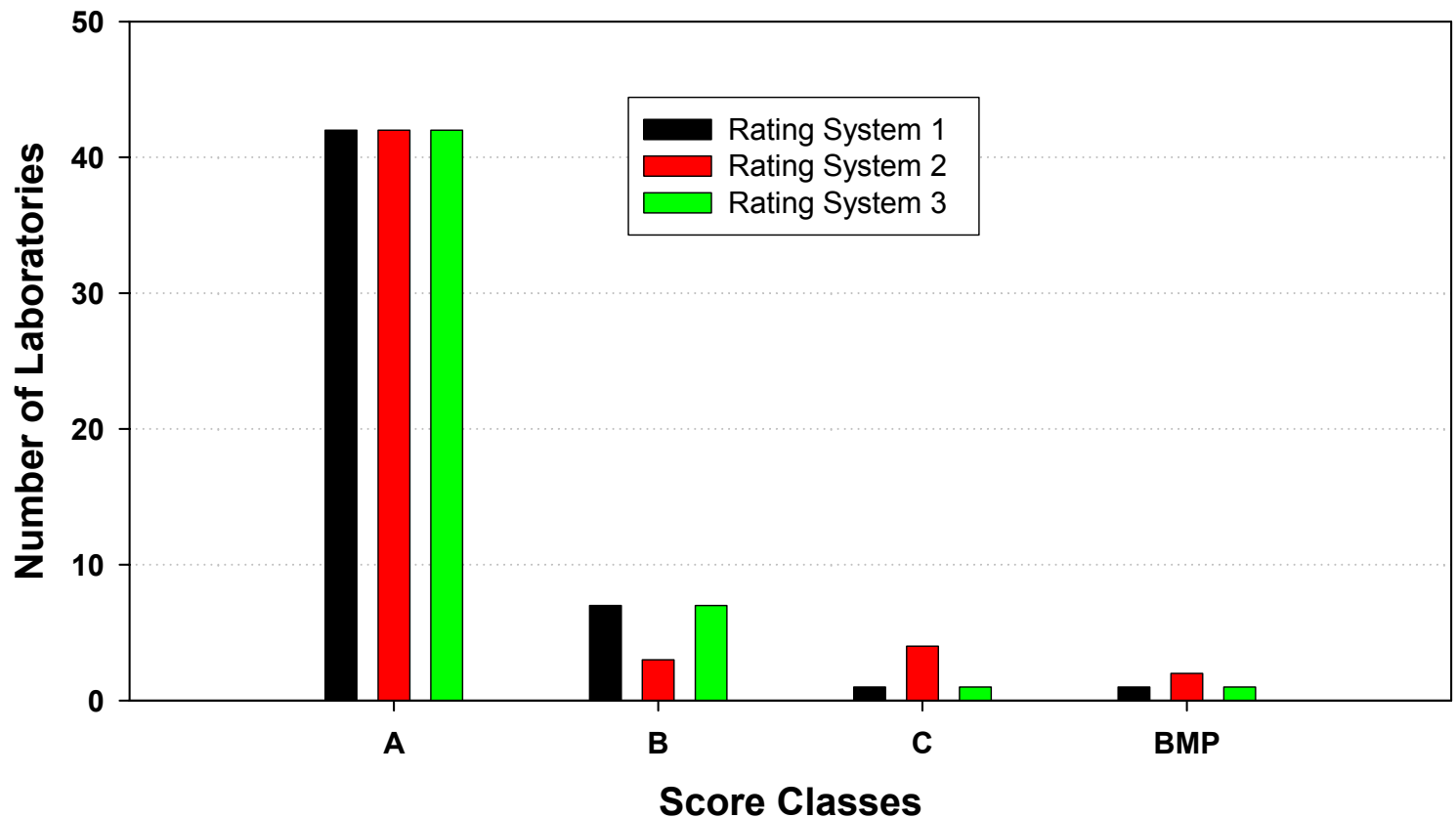
Rate	Fixed Number of misclassified samples
A	0 errors
B	1 or 2 errors
C	3 errors
BMP	More than 3 errors

Rate	Percentage of misclassified samples
A	0 or 1 error
B	2 errors
C	3 or 4 errors
BMP	More than 4 errors

Rate	Percentage of misclassified samples
A	0 or 1 error
B	2 to 4 errors
C	5 or 6 errors
BMP	More than 6 errors


 System 1 and system 2
 are about the same

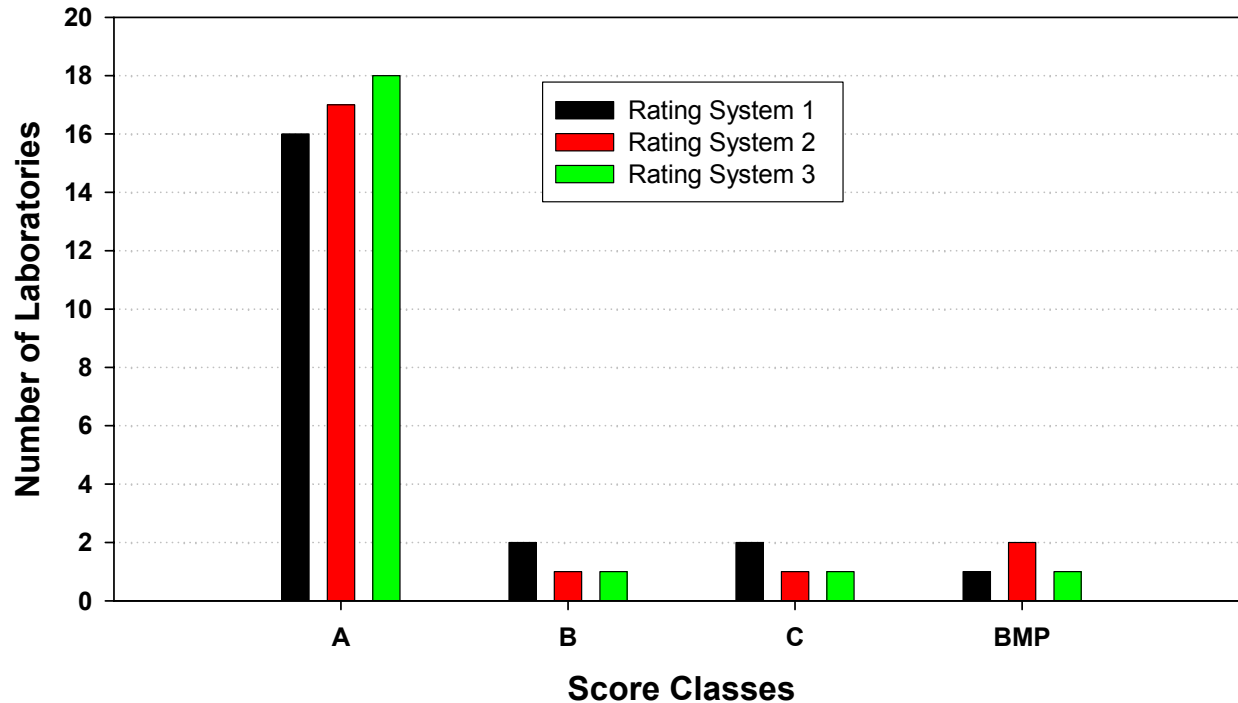
Comparison of the 3 Rating Systems for the Laboratories participated in ISTA Proficiency Test 4 only



Presence/absence rating systems

21 laboratories PT01+PT02+PT03

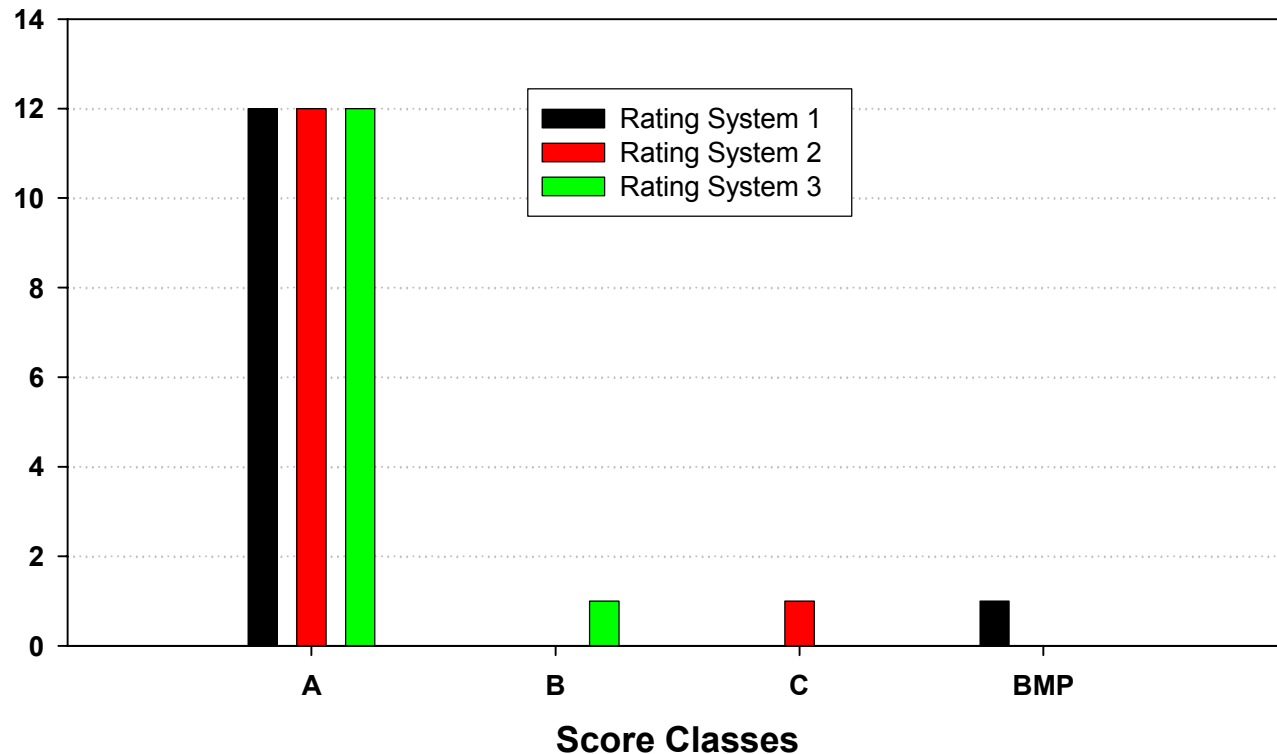
Comparison of the 3 Rating Systems for the Laboratories participated in the 3 ISTA Proficiency Tests



Presence/absence rating systems

13 laboratories PT01+PT02+PT03+PT04

**Comparison of the 3 Rating Systems for the Laboratories
participated in all 4 ISTA Proficiency Tests**



Quantification

3 suggested rating systems

- ❑ System 1: true value is number of GM seeds/total number of seeds in each sample
- ❑ System 2: true value is weight of GM seeds/total weight of seeds in each sample
- ❑ System 3: true value is the median of values reported by participating laboratories (after Cochran test)

There is no link between [system1 for presence/absence] and [system1 for quantification]

Computations are performed from all sample results returned by the laboratories whatever the method (including sub-sampling and quantitative PCR)

Where does these rating criteria come from?

It has been worked out for years (as soon as GMOTF was established) Joint group of STA committee and GMO Task Force

- ❑ In line with strategy position paper
- ❑ Adapted to the successive versions of Chapter 8 rules
- ❑ In line with ISTA proficiency test system

- ❑ Different statistical options were developed and compared (pooling, Bayesian approach, estimation and robustness of error rates,...)
- ❑ Checked for adequacy to data from actual situations
- ❑ Checked for consistency with more sophisticated analysis (ie Mixed models)
- ❑ Checked for Consistency with testing plan design in routine (ie Seedcalc)

Selection of the criteria, among appropriate ones

transparency,

« easiness » to compute and understand (Excel, BMP=pencil)

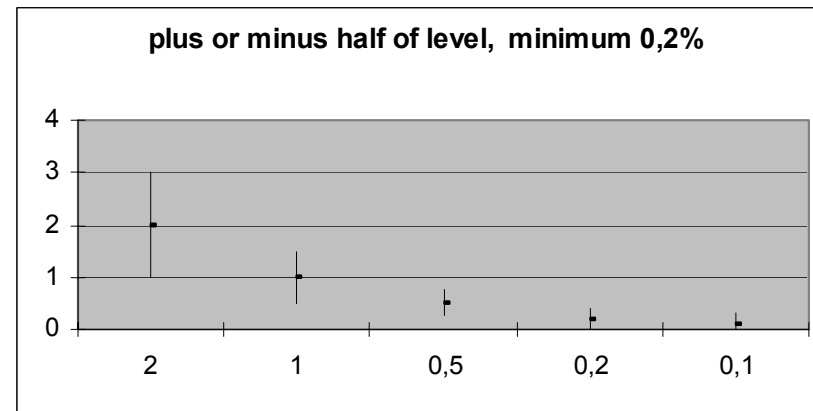
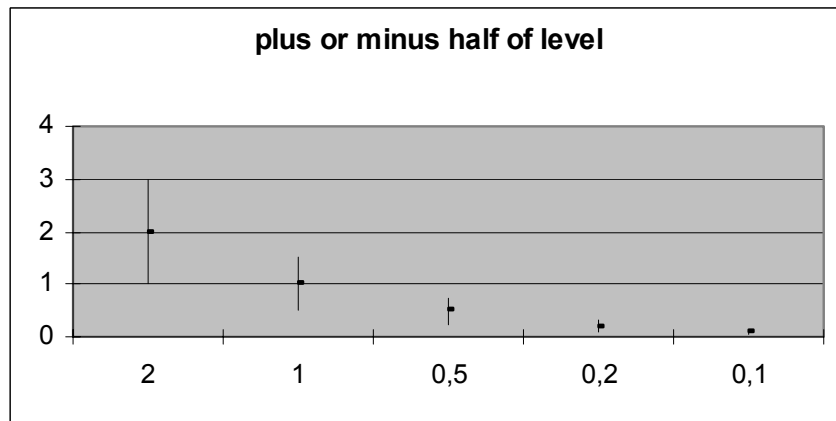
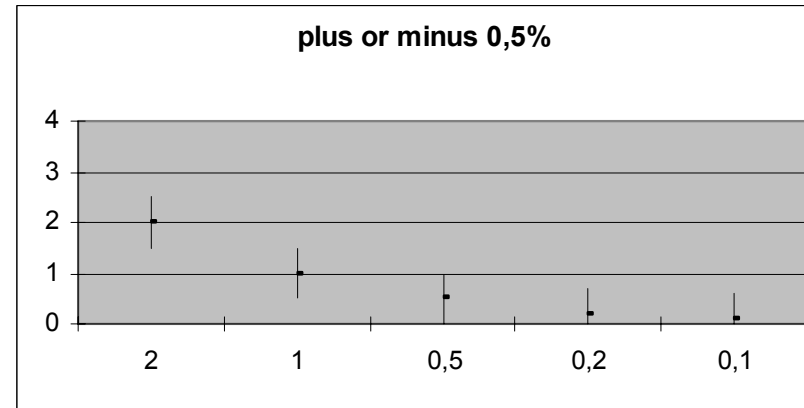
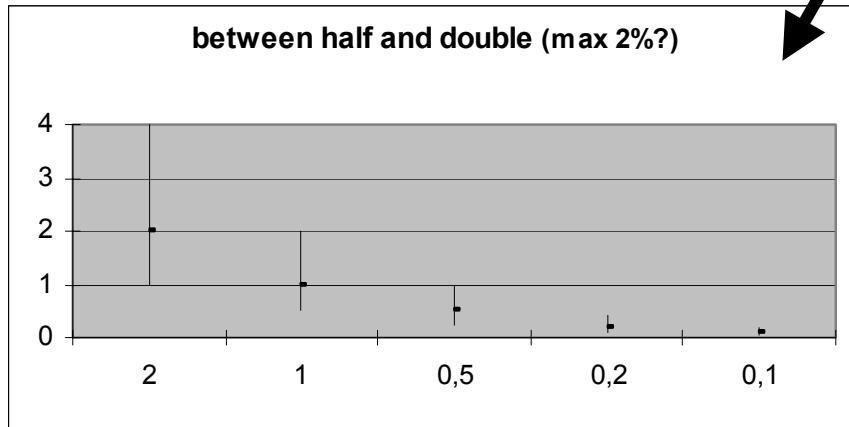
Rating principles for quantification

- *True value is known (purity check + event + number+ weight), and/or can be estimated from all results received*
- Too many *sample results* are too far from truth
->BMP
- *Average results* by spiking level are not accurate
->C
- Too many inaccurate *sample results* are not accurate
->B
- Otherwise rate A

NB: Computations are made on GM spiked samples

BMP: too many sample results (many could be $\frac{1}{4}$ $\frac{1}{3}$ $\frac{1}{2}$ $\frac{2}{3}$ of data points) are too far from truth (4 possible criteria are shown below)

Selected criterium: More than half of the results are outside half and double of truth



Within vertical bar = not too far from truth

C: The average results (from 2-6 samples) by spiking level are not accurate

- Mean of sample results per spiking level is computed
 - A « *spiking level z_score* » is computed for each spiking level.
 - A sum of absolute « *spiking level z_scores* » for the different spiking levels is compared to a statistical threshold
- This is comparable with C rating used in Germination rating for instance*

Normal Seedlings			Abnormal Seedlings			Non-germinated Seeds		
S1	S2	S3	S1	S2	S3	S1	S2	S3
88.9	92.7	88.4	5.02	3.07	6.98	5.04	3.68	4.33
90.3	89.5	88.8	6.75	9.25	7.50	3.00	1.25	3.75
0.4	-1.5	0.1	0.8	5.3	0.2	-1.6	-1.9	-0.4

✓ Sum of absolute Z-scores for one component > 5.3 ⇒ C-Rating

✓ Number of out of tolerance results: 1 ⇒ A-Rating

⇒ **In round: C**

5.3 for 3 values
(germination)

5.25 for 3 spiking levels
(GMO)

6.43 for 4 levels

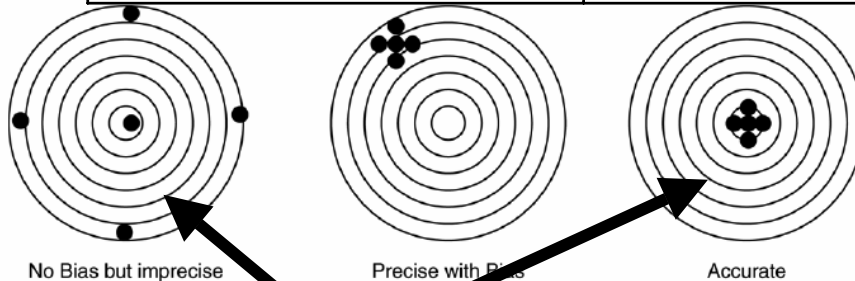
Number of spiking levels in the proficiency test	0.99 quantile
1	2,55
2	3,97
3	5,25
4	6,43
5	7,55
6	8,67
7	9,75
8	10,81
9	11,82
10	12,90

B: There are too many inaccurate sample results

- Individual « *sample z_scores* » are computed
- The number of « big » absolute « *sample z_scores* » is counted
- If number of big « *sample z_scores* » exceeds a pre-defined limit then B rating, otherwise A rating

Number of <i>non zero level</i> samples in the proficiency test	Maximum tolerated number of <i>sample z_scores</i> out of the interval [-2,+2]
1 to 5	0
6 to 11	1
12 to 17	2
18 to 23	3
24 to 29	4

More than 1/6th are inaccurate => rate B



Accurate =
Not far from true value
and
Not too much variable

5 samples in a spiking level : not looked at by C rating checked by B rating

A: No problem was detected

BMP check passed



C check passed



B check passed

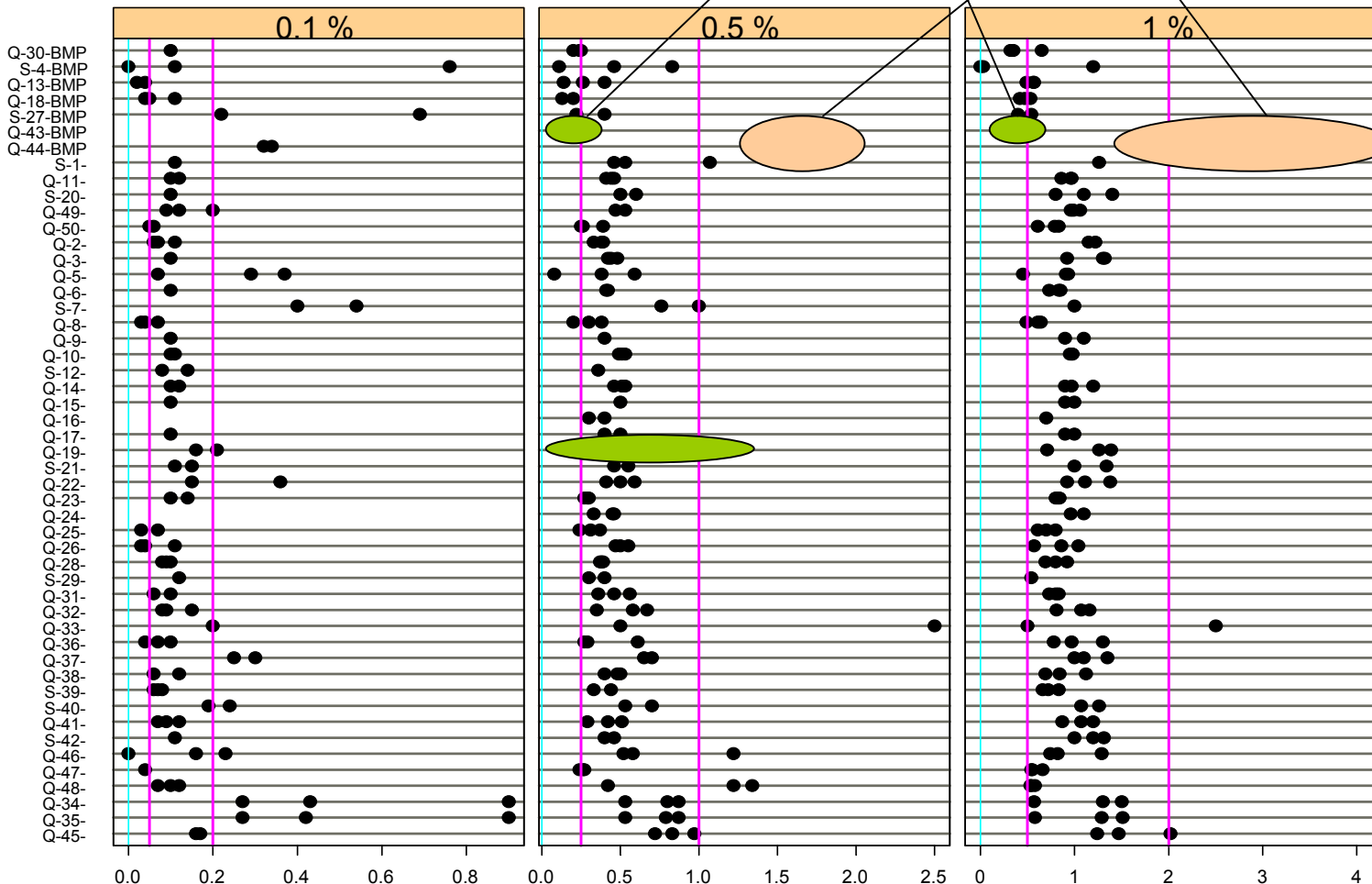


=>A rating.

Graphical outputs

1 line = 1 laboratory

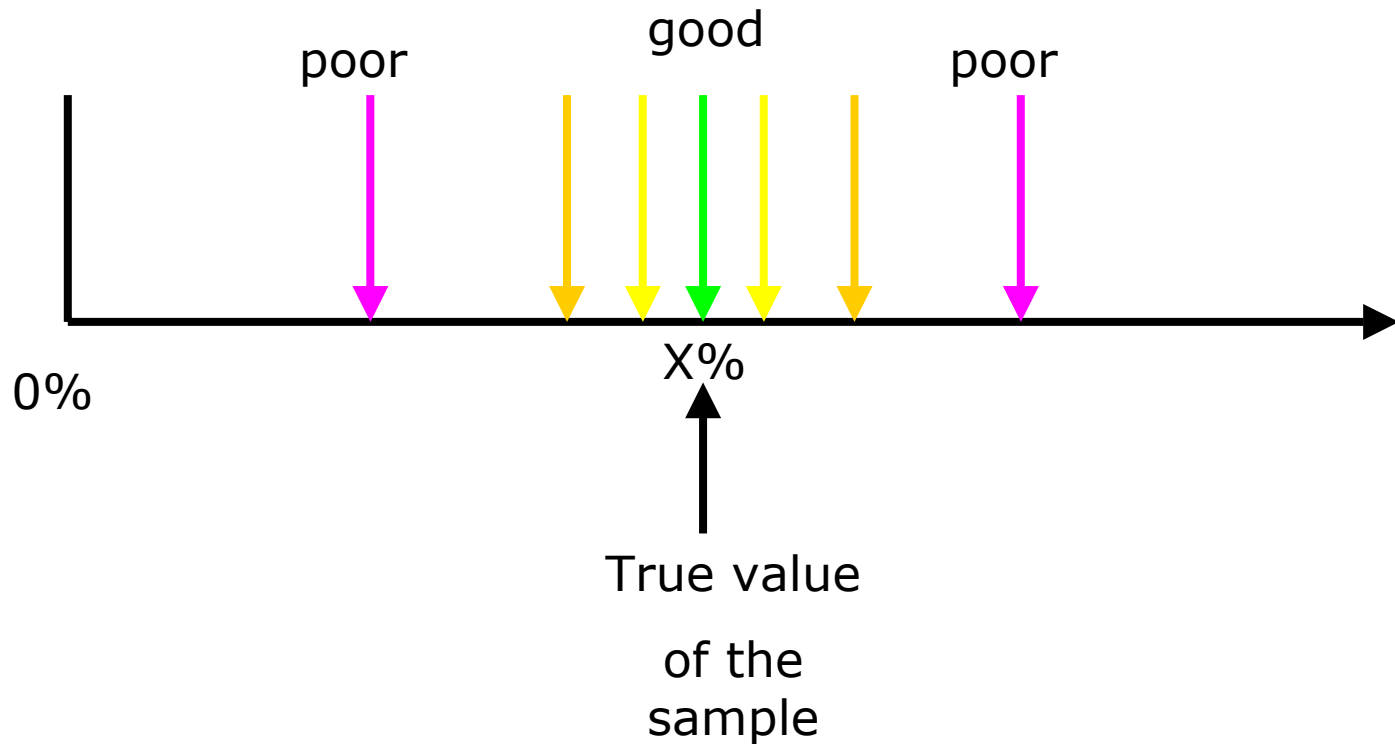
Reference: True level in % seed



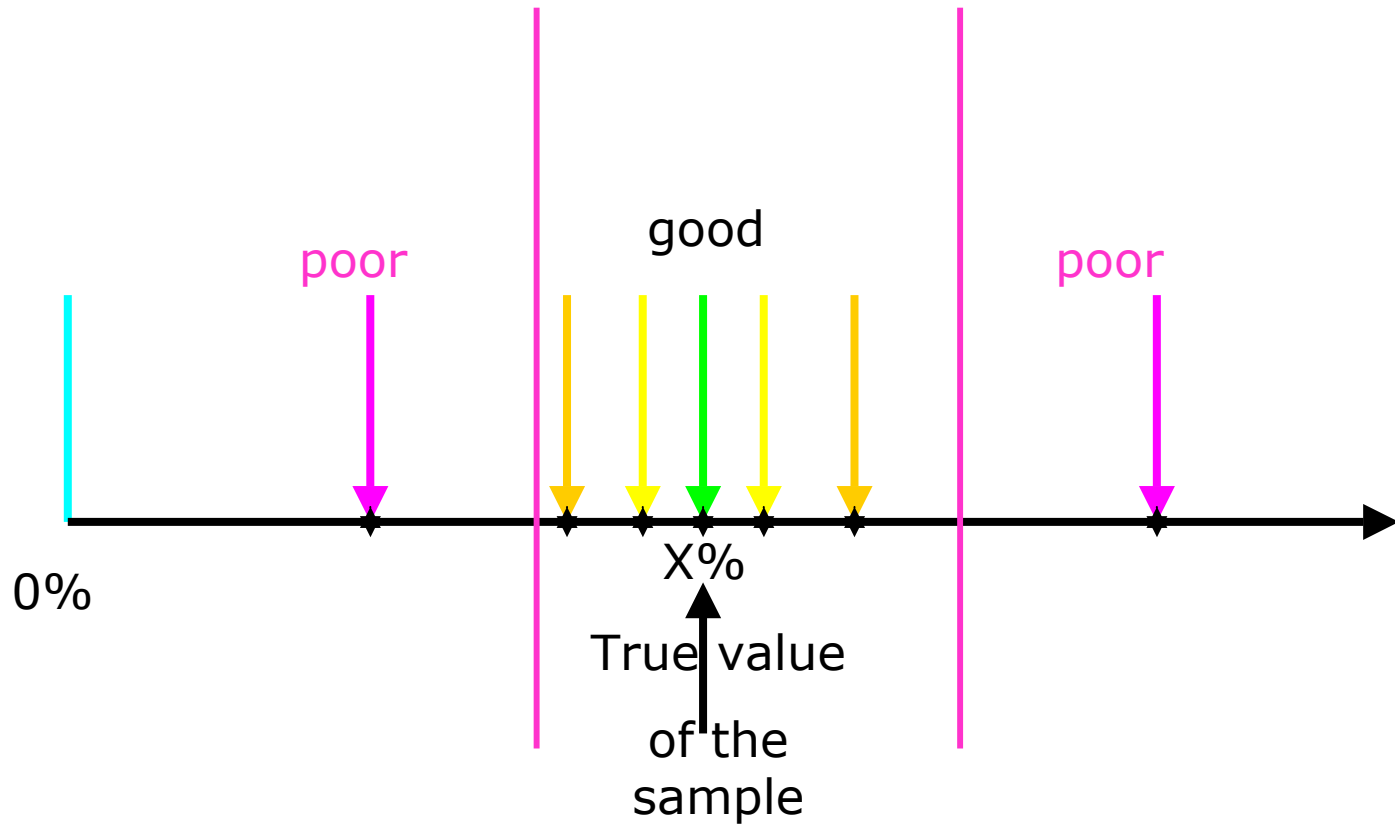
BMP Check example

Results

quality of a result on 1 sample



Limits shall be defined



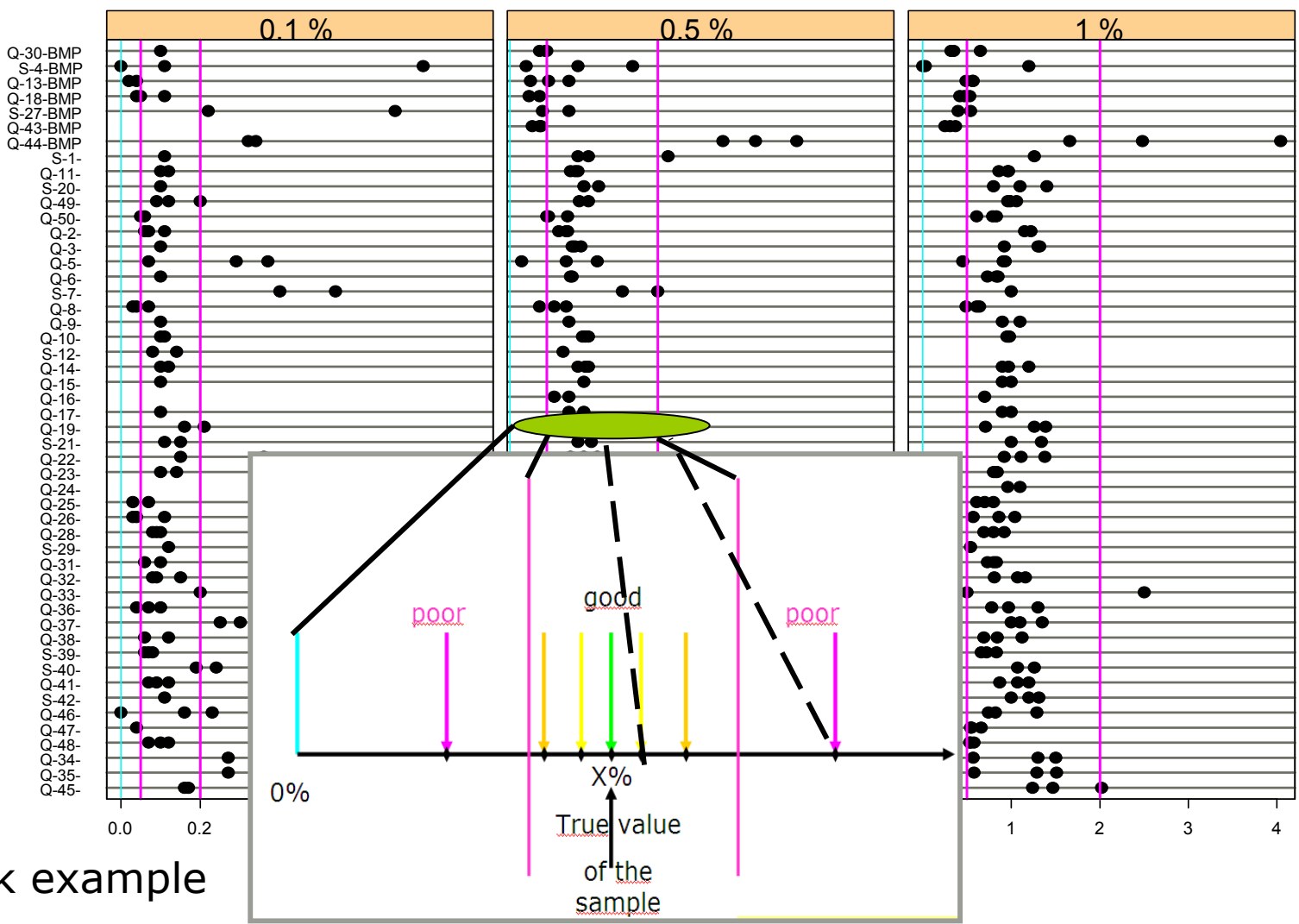
Reported value - true value

Reported value - true value
Common variability

Both options are used

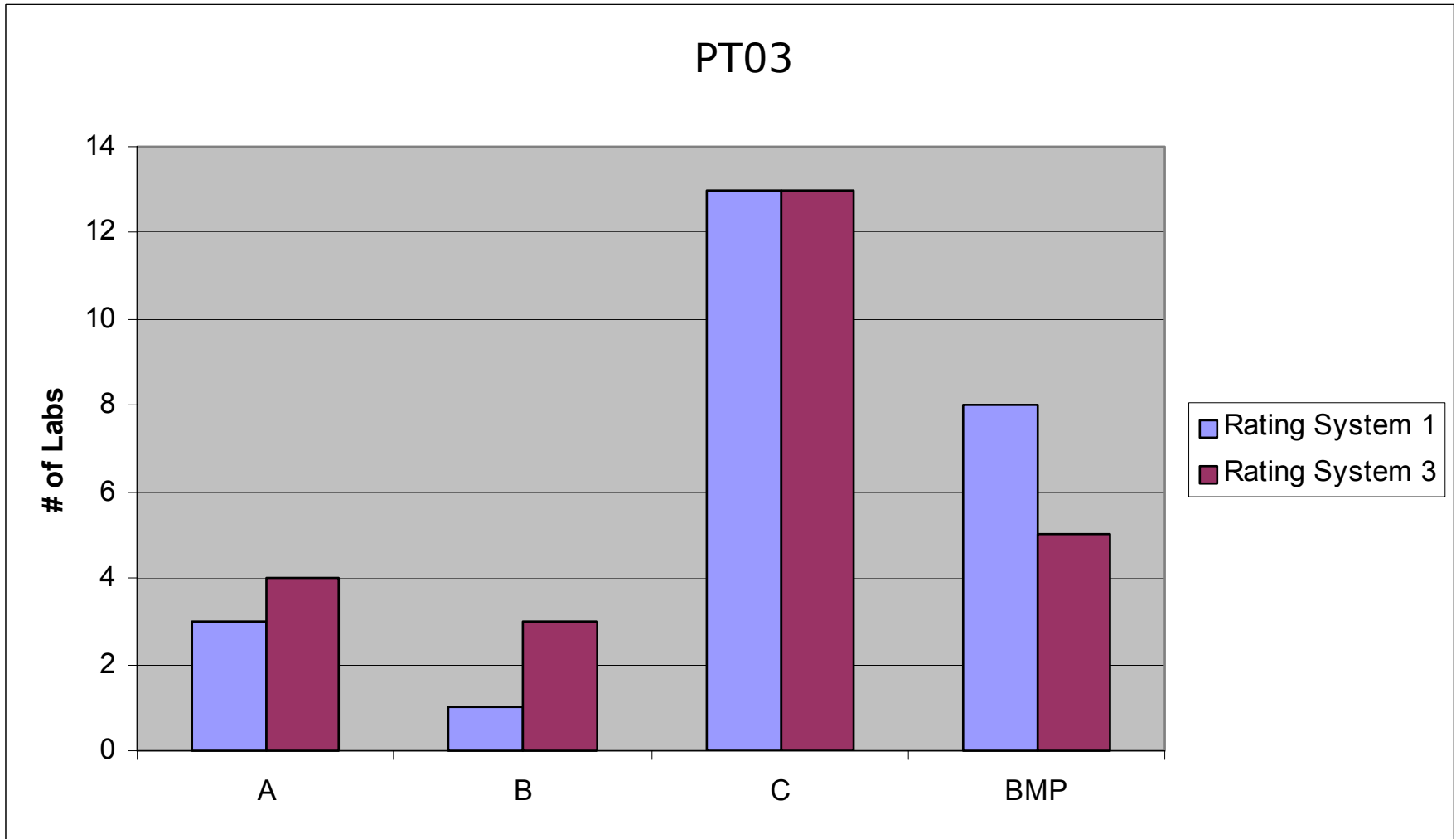
Same types of graphs are available for BMP, C, B and A rating, with their specific thresholds

Reference: True level in % seed



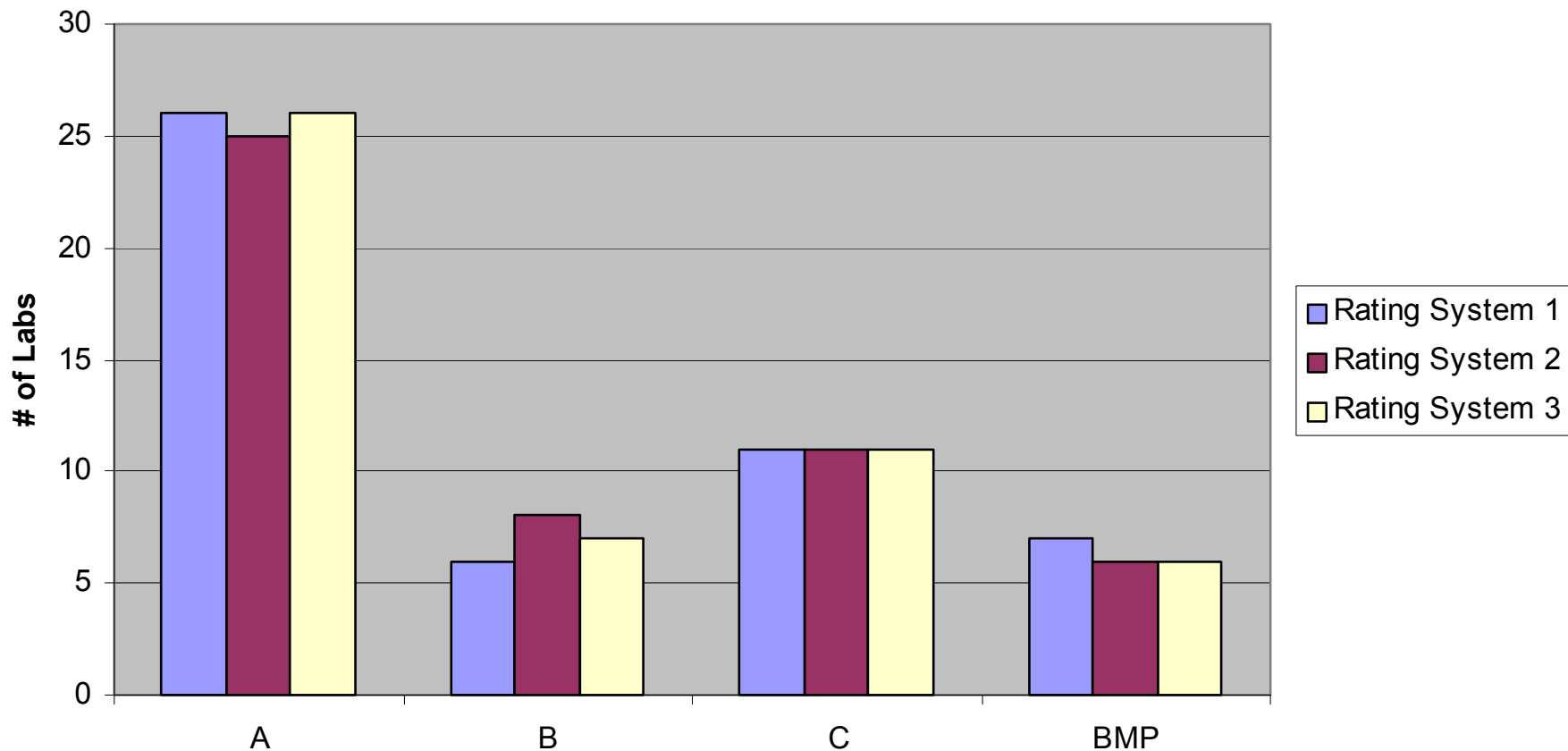
BMP Check example

PT3 rating summary for quantitative and semi-quantitative tests 25 laboratories



PT4 rating summary for quantitative and semi-quantitative tests 50 laboratories

PT04



Summary on suggested rating systems

- ❑ Follow same ISTA principles as for other types of tests
- ❑ Consistent with computational ratings already used in other types of tests
- ❑ Statistically appropriate for this specific type of test (presence, quantification)
- ❑ Appliance (ad hoc, fair and robust) checked with non ISTA proficiency tests, and data from companies/laboratories
- ❑ Backed up by more sophisticated statistical models, and careful look by different types of experts



Use of the detailed results provided by the Labs for the Quantitative Test (flour sub-sample results, measurement – replicates - results)

The dataset corresponding to the results from a particular laboratory is analyzed using a heteroscedastic linear mixed effects model:

$$Y_{ijkl} = \mu_i + A_{j(i)} + B_{k(ij)} + E_{ijkl} \quad (1)$$

where:

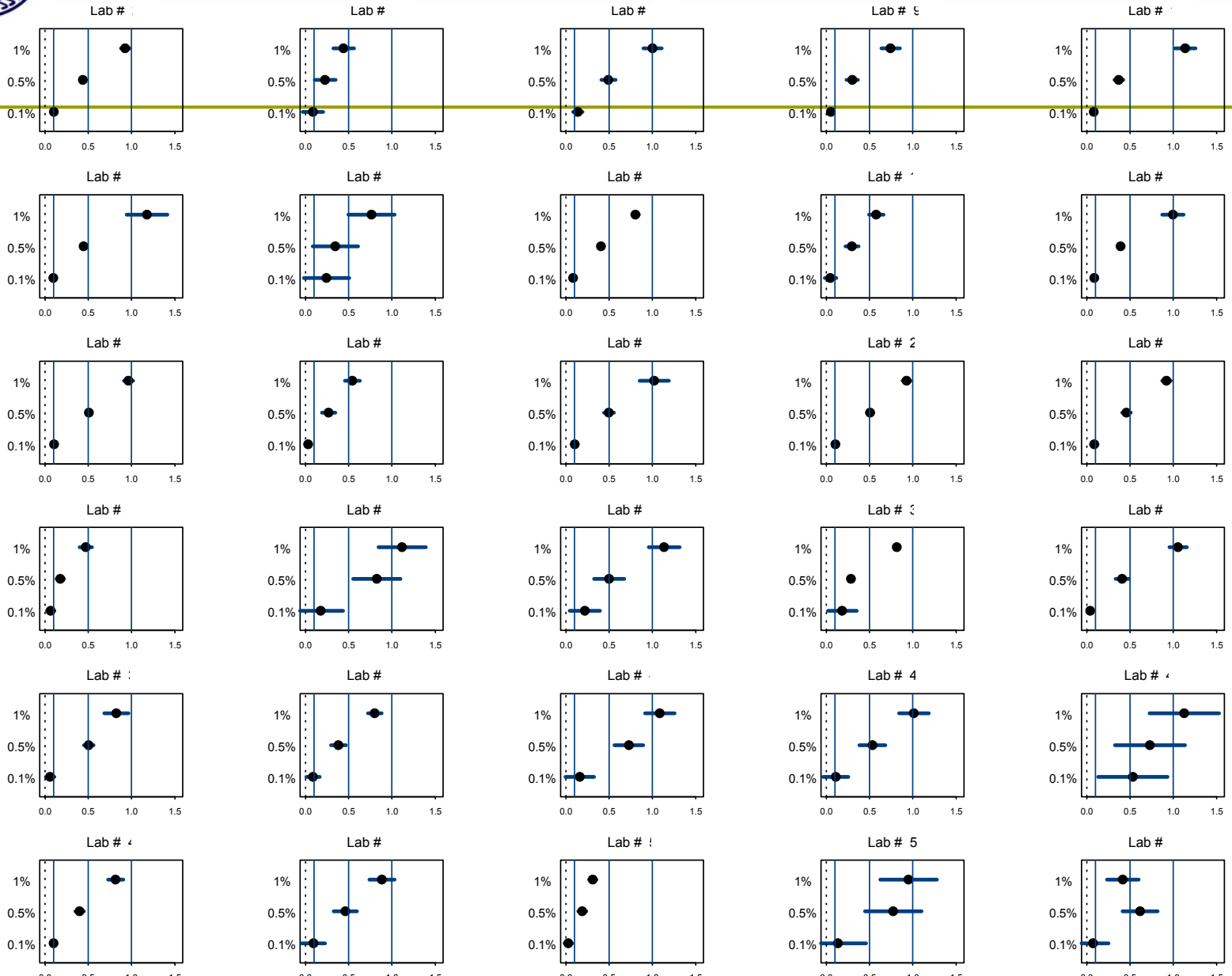
- μ_i is the mean for the i^{th} spiking level. Only spiking levels 0.1%, 0.5% and 1% are retained for this analysis ($i= 1, 2, 3$).
- $A_{j(i)}$ is the random effect of the j^{th} sample ($j= 1, 2, 3$) with spiking level i . The $A_{j(i)}$ are i.i.d. $N(0, \sigma^2_{\text{sample}})$, where i.i.d. is used to indicate that the observations are independently and identically distributed.
- $B_{k(ij)}$ is the random effect of the k^{th} flour sub-sample from sample j and spiking level i . The $B_{k(ij)}$ are i.i.d. $N(0, \sigma^2_{\text{flour}})$.
- E_{ijkl} are the measurement errors:

$$\left\{ \begin{array}{l} E_{1jkl} \text{ are i.i.d. } N(0, \sigma_1^2) \\ E_{2jkl} \text{ are i.i.d. } N(0, \sigma_2^2) \\ E_{3jkl} \text{ are i.i.d. } N(0, \sigma_3^2) \\ \text{cov}(E_{ijkl}, E_{i'j'k'l'}) = 0 \text{ for } i \text{ different from } i' \end{array} \right.$$

Mixed-effects model

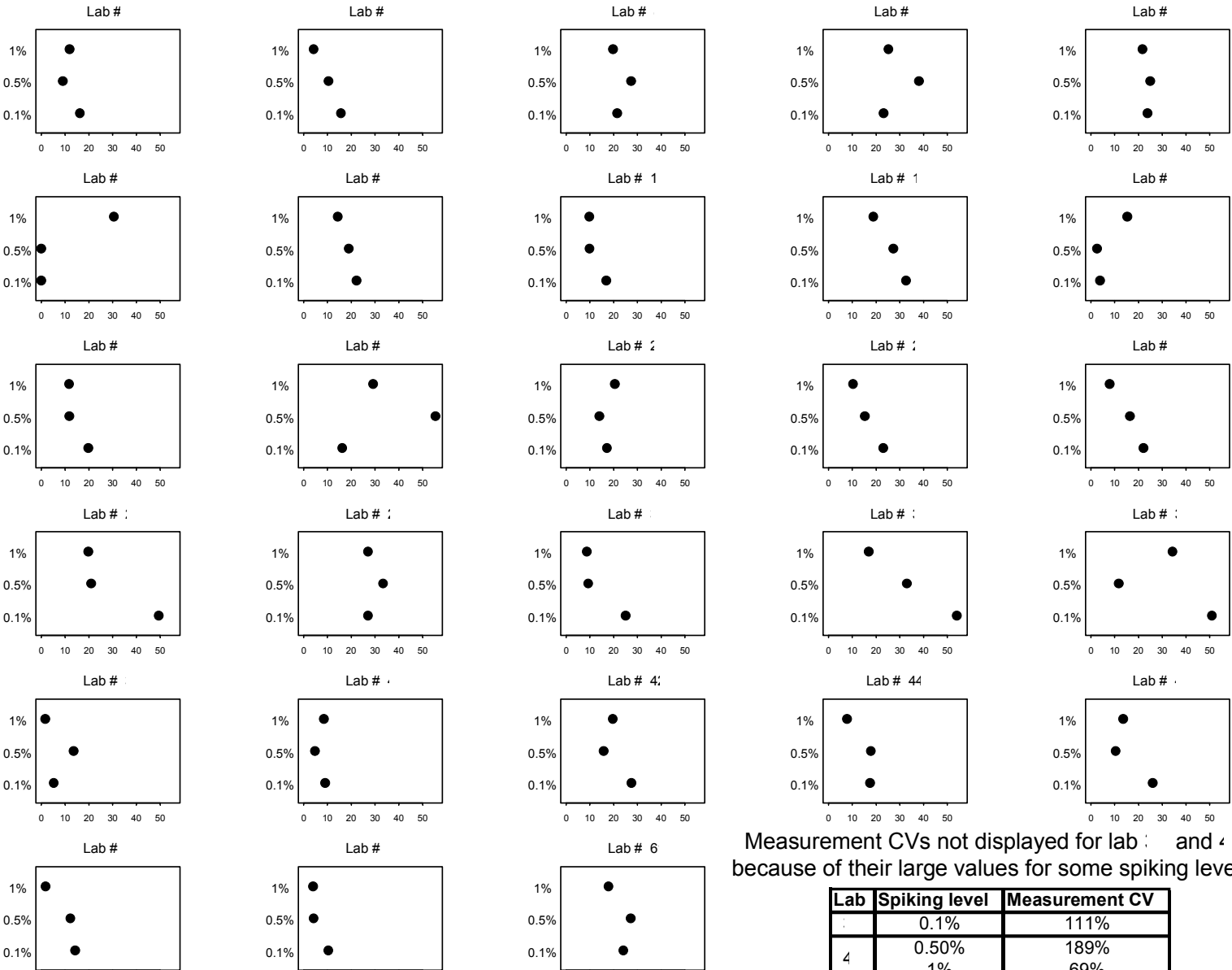


For each lab and each spiking level, mean and its associated 95% confidence interval





For each lab and each spiking level, measurement CVs



Measurement CVs not displayed for lab : and 4 because of their large values for some spiking levels:

Lab	Spiking level	Measurement CV
:	0.1%	111%
4	0.50%	189%
	1%	69%