ACTA IMEKO
ISSN: 2221-870X
September 2023, Volume 12, Number 3, 1 - 9

β-risk in proficiency testing in relation to the number of participants

Louis-Jean Hollebecq[1]

[1] *CompaLab, 16 avenue du Général de Gaulle, 93110 Rosny-sous-Bois, France*

ABSTRACT
The Monte Carlo method was used to investigate the capacity of PT schemes to detect laboratories which biases are out of control. Probabilities that the computed z values are over 3 while the true value is less than 2 and that the computed z values are less than 2 while the true values are over 3 are computed for a series of situations: number of participants from 5 to 30, various ratios of repeatability over reproducibility and number of test results per participant, introduction or not of outliers with *k* from 3.5 to 10. For each situation, the probabilities of not detecting true outliers and to trigger false alerts are discussed. Guidance and keys are proposed to check and improve the efficiency of real PT programs.

Section: RESEARCH PAPER

Keywords: Laboratory proficiency testing; Monte Carlo methods; efficiency of assessment

Citation: Louis-Jean Hollebecq, β-risk in proficiency testing in relation to the number of participants, Acta IMEKO, vol. 12, no. 3, article 11, September 2023, identifier: IMEKO-ACTA-12 (2023)-03-11

Section Editor: Marija Cundeva-Blajer, Ss. Cyril and Methodius University in Skopje, North Macedonia

Received December 20, 2022; **In final form** May 15, 2023; **Published** September 2023

Corresponding author: Louis-Jean Hollebecq, e-mail: ljh@compalab.org

1. INTRODUCTION

Proficiency tests (PT) are widely used to assess the performance of laboratories. Participating to such programs is required by ISO 17025 [1], which is the standard of reference for accreditation of laboratories.

Reference standards for interlaboratory comparisons (ILC), ISO 5725-2 [2], ISO 13528 [3] and ISO 17043 [4] consider one or several aspects of the performance of laboratories:

1. Assessment of the bias of participants. D, D% and z-scores (or equivalents) are related to this assessment;
2. Assessment of the uncertainties on test results claimed by participants. ζ-scores and E_n (or equivalents) that are related to this assessment;
3. Assessment of repeatability. Some methods that are presented as "graphical" in [3], are related to this assessment.

These 3 types of assessment are totally different in nature: assessing a bias has something to do with assessing a mean value, while assessing uncertainties or repeatability has something to do with assessing a standard deviation. Consequently, totally different studies need to be conducted to address these issues. The scope of this study was limited to the assessment of biases. Further studies are needed for the other types of evaluation.

Moreover, in the case of the assessment of the bias, both the assigned value and the maximum acceptable deviation can be determined by several means. [3] provides an extensive discussion concerning advantages and disadvantages of several of them. We decided to limit our study to the use of z-scores computed from the results of participants for the following reasons:

1. z-scores (or equivalents) are the most used in practice (in many cases, other methods cannot be applied for technical or practical reasons);
2. When the assigned value and/or acceptable deviation are externally determined, several additional parameters (typically, the difference between the assigned value and the "true" average value of results, and the difference between the acceptable deviation and the actual scatter of test results of participants) need to be considered. On the other hand, no effect of estimation of these parameters from the test results of participants happens, what lowers the added value of the use of the Monte-Carlo method, because it is precisely on this point that the use of this method makes sense.

[2], [3] and [4] define limits for computing the alerts with regard to z-scores, corresponding to theoretical risks of 5 % and 0.3 %.

Note that [2] deals with ILC to assess test methods. [3] deals with ILC for PT of labs. [4] is the reference for accreditation of PT providers. Consequently, even if the theoretical risks in all of them are same, their goal is different, in accordance with the aim of the standards:

3. Limits in [2] are intended to assure the reliability of the assessment;
4. Limits in [3] and [4] are for proficiency checking of participants.

[2] is referred here because it is the "historical one" and it is still widely used by PT providers, even if [3] is obviously better adapted to proficiency testing of labs.

These risks are of α-type (risk to trigger a warning that should not). Another risk actually occurs, usually called β-type (risk of not triggering a warning when it should). However, even if this question is of main importance, this β-type risk is quite hard to compute, and for this reason, is almost always just ignored, including in the reference standards [2] and [3]. Everybody knows that an enough number of participants is necessary to ensure the efficiency of the PT, but there is no clear consensus of what that "enough number" should be. On the other hand, test methods for which there are very few potential participants to a PT are quite numerous. There is then no opportunity for these laboratories to get the advantages of a participation to a PT. This paper proposes to overcome the difficulty of computing the β-risk by using the Monte-Carlo method and to provide a beginning of answer to the question: does it make sense or not to organise PTs with 5 or 8 or 12 participants, especially when the number of potential participants is quite low?

To do so, the following issues are dealt with:

1. How α-type and β-type risks can be computed and what hypotheses to do it were taken into account in the present study;
2. What are the principles of the Monte-Carlo method, in which conditions it can be used and how it was implemented in the present study;
3. What is the impact of the use of robust statistics that are usually used to avoid the deleterious impact of outlying results on the so-called assigned values;
4. What is the impact of the number of test results by each participant, with regard to interlaboratory and repeatability standard deviations.

2. DESIGN OF EXPERIMENTS

2.1. List of symbols

The symbols used in this article are listed in Table 1.

2.2. Calculation of α-type and β-type risks

Computing α-type and β-type risks requests to define underlying alternate hypotheses usually designated as H_0 and H_1. α is the probability to reject the H_0 hypothesis while it is actually true and β is the probability to reject the H_1 hypothesis while it is actually true, as shown in Table 2.

The issue of α-type and β-type risks have been extensively discussed for a very long time because they address many practical decision problems, notably the assessment of conformity of products to specifications, see for example ISO 3951-1 [5]. In all cases:

Table 1. List of symbols.

Symbol	Designation
B	Bias of the lab in the model of ISO 5725-1
D and D%	Participant's difference with assigned value, absolute or in %, as defined in ISO 13528
e	Random error in the model of ISO 5725-1
E_n	Participant's score used to assess its uncertainties, as defined in ISO 13528
m	General mean value in the model of ISO 5725-1
N_r	Number of test results per participant
N_p	Number of participants
N_s	Number of samples distributed to each participant
s^*	s_{pt} computed with a robust algorithm (see ISO 13528)
s_{pt}	Estimate of σ_{pt} computed from the data of the PT
x_i	Result of participant "i"
X_{pt}	Central value or assigned value, that is used as reference value for the PT
z	Normalised participant's score used to assess its bias, as defined in ISO 13528, see Equation (2)
z_{true}	z-score that would be attributed to a participant if X_{pt} and σ_{pt} were exactly known
z_{calc}	z-score attributed to a participant, computed with X_{pt} and σ_{pt} determined from the test results of the PT
α	Probability to trigger a false alert for a participant
β	Probability not to trigger an alert for a participant when it should
ζ-score	Normalised participant's score used to assess its uncertainties, as defined in ISO 13528
λ	Parameter as defined in Equation (4) encompassing effects of σ_{rPT}, σ_{iL} and N_r on the efficiency of the PT scheme
σ_{BL}	Standard deviation of the biases of the participating labs
σ_H	Standard deviation representing the homogeneity of samples
σ_{iL}	Standard deviation due to internal scatter of the laboratory results other than repeatability (differences between operators, machines of the lab, variations of environmental conditions within the lab along the time)
σ_{pt}	Standard deviation assigned for the PT
σ_r	Standard deviation of repeatability, as defined in ISO 5725-1
σ_{rPT}	Standard deviation of sets of results of participants
σ_R	Standard deviation of reproducibility

1. α and β-risks decrease when the available number of test results increases;
2. For a given number of test results, α-risk increases when β-risk decreases, and vice-versa.

In the context of PT organisation, the H_0 hypothesis can be quite obviously defined as "The results of the participant belong to the general population of expected results". In the same way, H_1 can be defined as "The results of the participant belong to a population other than the one of the expected results".

It is needed then to define how conclusions about H_0 and H_1 shall be carried out. The decision rules described in the reference standards [2] and [3], i.e. the calculations of z-scores obviously apply to H_0. On the contrary, the distribution of H_1 is not known (other populations of results than the expected one can

Table 2. α- and β-risks with regard to H_0 and H_1 hypotheses.

	H_0 is true	**H_1 is true**
H_0 is accepted	Right decision ($p = 1 - \alpha$)	Wrong decision ($p = \beta$)
H_1 is accepted	Wrong decision ($p = \alpha$)	Right decision ($p = 1 - \beta$)

practically be very different ones, including gross errors, different types of deviations to the method, etc. …). One way to solve this problem is to construct "power curves" in function of parameters of the problem, and especially the number of results and the distance to H_0. This principle was used to build up the design of experiments for this study.

In details:

1. We considered that the α-risk has occurred when $|z_{\mathrm{calc}}| > 3$ and $|z_{\mathrm{true}}| < 2$ (as recommended in [2] and [3]), and that the β-risk has occurred when $|z_{\mathrm{calc}}| < 2$ and $|z_{\mathrm{true}}| > 3$;
2. We computed these α and β-risks on populations of test results without any true outlier, i.e. to a whole Gaussian population of expected results. This implicitly includes 5 % of corresponding z-scores outside the [-2;+2] interval and 0.3 % outside the [-3;+3] interval;
3. We also computed the α and β-risks on populations of test results including one true outlier with various z values from 3.5 to 10. These computations of α and β-risks were carried out separately for the main population and for the outlier, enabling to check the impact of the outlier on both categories of participant results.

It should be kept in mind that the computed β-risks fully depend on the definition of H_1 (see here upper) and that other ways to define H_1 would also make sense, leading to other meaningful values of β.

To deal with the upper, we have built up a design of experiments pursuing the following goals:

1. Impact of the number of participants;
2. Principles of the Monte-Carlo method;
3. Impact of the type of statistics used to compute the so-called assigned values;
4. Impact of the number of repetitions by each participant, with regard to interlaboratory and repeatability standard deviations.

Each of these issues are developed here after.

2.3. Impact of the number of participants

[1] recommends that at least 12 participants are present and [2] recommends not to use robust statistics when the number of participants is less than 18. On the other hand, our computations showed that α and β-risks do not significantly change when the number of participants goes over 30. We then did not investigate higher values and decided to compute the α and β-risks for a number of participants varying from 5 to 30. This enabled us to investigate areas that are not recommended by the standards and compare them to recommended ones.

2.4. Principles of the Monte-Carlo method

The Monte-Carlo methods are a large category of algorithms that use random numerical realisations of a given model. They are often used to solve mathematical or physical problems, difficult or impossible to solve by other methods. For a survey of the history and applications of the Monte-Carlo methods, see for example [6].

In our problem, using the Monte-Carlo methods enables us to create series of "true values" of test results that cannot be known in real life. In practice, we always know whether H_0 and H_1 are accepted or not (i.e. whether an alert was sent to the participant or not), but we can never know whether H_0 and H_1 are actually true or not. Using Monte-Carlo methods enables us to control at the same time for each series of random results whether H_0 and H_1 are accepted or not and whether H_0 and H_1 are true or not. Having this whole information is necessary to compute both α- and β-risks.

However, using Monte-Carlo methods requests to use a model that reasonably fits the situations encountered in the real world. In this study, we used the model of ISO 5725-1 [7] widely used to cope with problems of precision of test results:

$$y = m + B + e\,, \tag{1}$$

where m is the general mean value, B is the bias of the lab and/or the method, and e is the random error.

In this model, we used $m = 0$, a Gaussian distribution with 0 as mean value and 1 as standard deviation for B and another Gaussian distribution with 0 as mean value and a varying σ_{r} as standard deviation for e (see at 2.6 a discussion about what σ_{r} really represents in practice, with regard to the design of the PT scheme as decided by the PT provider).

Using the Monte-Carlo methods also requests to use random input values. When several random values are necessary to produce one Monte-Carlo result and when correlations between them apply in real life, these correlations must be incorporated in the input values of the computations. That can be a bit difficult to be done properly. In our case, the Monte-Carlo results are z-scores, and $N_{\mathrm{r}} \times N_{\mathrm{p}}$ random values (N_{r}: number of test results per participants and N_{p}: number of participants) are needed to compute them. We can reasonably rule out the existence of any correlation, assuming that there is no correlation between the results of the different participants and between the results of a same participant. As a matter of fact, [3] requests PT providers to care about that (no collusion between participants), because it is a condition to ensure the validity of the statistical treatment.

To assure the validity of the conclusions, the random series need to be numerous enough, depending on many factors. In our study, we computed series of 500 000 to 4 000 000 z-scores for each situation (i.e. for each combination of number of participants, number of test results per participant and $\sigma_{\mathrm{rPT}}/\sigma_{i\mathrm{L}}$ ratio). Each of them was divided in 40 sub-groups enabling us to check how repeatable were the computed α and β within the 40 sub-groups and compute a related interval of confidence (IC). This IC always happened to be always less than $\pm$ 2 % (with enlarging coefficient $k = 2$) and in all cases significantly lower than of the computed α and β.

2.5. Impact of the type of statistics used to compute the so-called assigned values

Results with gross errors often occur during the organisation of PT. They are usually caused by typing errors, by misunderstanding of instructions for participation or by using wrong units. In most cases, gross errors are due to necessary deviations to routine procedures of the labs when they participate to a PT. Typically, typing errors usually never occur in real life because data transfer is nowadays never performed manually, contrarily to the cases of participations to PT.

However, gross errors are a big problem for the statistical data processing, because they strongly impact the estimation of statistical parameters, making them irrelevant. In particular, they strongly increase the computed standard deviations, and hence, the β-risk. On the other hand, just ignoring the suspicious results might lead to underestimate the standard deviations of reference, increasing then the α-risk.

To face this problem, [2] and [3] recommend to detect outliers and/or use so-called robust statistics. These robust statistics generally consist in replacing outlying results by softened virtual

ones, using algorithms specifically designed for that. Full information about this can be found in particular in Annex C of [3] and in [8]. These robust methods tend to produce mean values and standard deviations resisting to a certain proportion of outliers (called breaking point) but also to decrease the speed of convergence of the estimates towards their central values. Annex D of [3] provides a comparison of the breaking points and speeds of convergence of the different algorithms that it proposes.

Because of the decrease of the speed of convergence in estimations of mean values and standard deviations, [3] and [9] recommend not to use robust statistics for a low number of participants. However, [10] and [11] went quite deeper in studying the issue and both showed that using robust statistics considerably improve the estimation of central value and scatter of the distribution in presence of outliers and consequently improve the assessment of the performance of PT with low number of participants. They both compared the different robust methods usually used but they both conclude that their relative efficiency depends on the type and proportion of outlying results.

On our own, as PT provider, our line of action has always been to use robust statistics, even for low number of participants, preferring running the risk of a day-to-day slightly lower efficiency of assessment than a risk of completely misleading one, even sporadically.

For the sake of this study (which is not to compare the efficiency of the different available robust methods), we chose to compute α and β-risks without robust statistics and with the so-called A algorithm described in [2] and [3], which is the most widely used by PT providers. This enables us to check the impact of using robust statistics or not without increasing to much the volume of needed calculations.

In order to check the impact of outliers, we produced series of test results without outliers and with one outlier which true z-score varies from $z = 3.5$ to $z = 10$. It follows that the proportion of outliers depends on the number of participants N_p, from 20 % for $N_\mathrm{p} = 5$ to 3.3 % for $N_\mathrm{p} = 30$. This option does not necessarily represent faithfully what happens in practice (see [10] and [11] for that), but we chose it because:

1. It does not request any modelling of outlying;
2. And it provides information about the impact of outliers easier to handle.

2.6. Impact of the number of repetitions by each participant, with regard to interlaboratory and repeatability standard deviations

In almost all cases, PT providers use z-scores or equivalents to assess the performance of the participants. According to [3] and [4], z-scores can be computed according to the equation (2):

$$z = \frac{x_i - X_\mathrm{pt}}{\sigma_\mathrm{pt}}, \tag{2}$$

where x_i is the result of the participant i,

X_pt is the central value

and σ_pt is the standard deviation assigned for the PT.

The performance is regarded as satisfactory when $z \in [-2; +2]$ and not satisfactory when $z \notin]-3; +3[$.

Note that these limits are completely conventional. They implicitly refer to the idea that the probabilities for these events to occur are respectively 95 % and 0.3 %, and other choices would also make sense. Consequently, the theoretical α-risk is 0.3 %. In other words, the probability to decide that the results are unsatisfactory, while in fact they do, belong to the main population is 0.3 %.

In fact, this would be true if σ_pt had exactly represented σ_BL, standard deviation of the biases of all the participating laboratories, what is never true. In most cases, σ_pt is computed as s_pt (or s^* when a robust algorithm is used), defined in [2] and [3] as the standard deviation of the results of all participants. Then, in practice, σ_pt can be computed with the equation (3):

$$\sigma_\mathrm{pt}^2 = \sigma_\mathrm{BL}^2 + \sigma_\mathrm{iL}^2 + \frac{\sigma_\mathrm{r}^2}{N_\mathrm{r}} + \frac{\sigma_\mathrm{H}^2}{N_\mathrm{s}}, \tag{3}$$

where σ_BL is the standard deviation of the biases of the participating laboratories,

σ_iL is the standard deviation due to internal scatter of the laboratory results other than repeatability (differences between operators, machines of the lab, variations of environmental conditions within the lab along the time),

σ_r is the repeatability standard deviation,

N_r is the number of test results per lab,

σ_H is the standard deviation representing the homogeneity of samples and

N_s is the number of samples provided to each lab.

In order to ensure the efficiency of the PT, PT organisers normally request the participants to produce their results in repeatability conditions, as defined in [2]. That is to say, the results provided by the participant are normally coming from a same operator using the same equipment, and tests being performed in a short period of time. However, in its day to day life, the laboratory usually produces test results from several operators, using different equipment in testing conditions that vary along the time. Consequently, the mean value of the test results that the participant delivers to the PT organiser randomly distributes around its yearly global mean value, with a standard deviation σ_iL representing the scatter due to the effects of using different operators, different equipment, and different test conditions along the time.

This standard deviation σ_iL is usually unknown because computing it is quite complicated. Indeed, to evaluate it properly, testing plans need to cover several operators, various equipment, and a long period of time. However, it is possible to compute it for example when PT are organised with a high frequency (for example once a month) or when the laboratory has put in place a surveillance of its test results along the time by using a control chart, provided that the corresponding results represent all the test conditions along the time. We did not consider it here because it hardly happens. When relevant (i.e. when σ_iL has technical reasons to be important), PT organisers could request participants to produce several series of results corresponding to different operators, equipment and testing conditions. ISO 5725-3 [12] describes efficient methods to determine intermediate fidelity and would be useful to perform that. In such cases, the term σ_iL^2 should then be transformed into $\sigma_\mathrm{iL}^2/N_\mathrm{iL}$ (where N_iL is the number of corresponding repetitions or the relevant number of degrees of freedom when "unbalanced testing schemes" are used).

On the other hand, the test results provided by the participant are made from a limited number of repetitions. Because of that, and when the participant actually produces its results in repeatability conditions, the mean value of the test results that the participant delivers to the PT organiser randomly distributes

around its mean value, with a standard deviation σ_r representing the repeatability. This effect is however softened by the number of repetitions, in accordance of the statistical law that applies for the estimation of a mean value, which justifies the contribution σ_r^2/N_r.

Except the cases where all participants perform their tests on the same samples (what can be done only if tests are not destructive and what generates practical difficulties of organisation), the samples on which the participants perform their tests can never be all exactly identical. Consequently, the mean value of the test results that the participant delivers to the PT organiser randomly distributes around its mean value, with a standard deviation σ_H representing the lack of homogeneity of distributed samples. This effect is however softened by the number of samples that are distributed to the participants. In the same way than for repeatability, the contribution of it is σ_H^2/N_s.

In addition to all of this, we have to stress out that all this is valid only if all the here upper described variances can be regarded as independent. If some correlations were to exist between those factors, then the corresponding covariances should be taken into account. This is obviously not the case in here: scatters due to interlaboratory effects, intra-laboratory effects, repeatability effects and homogeneity effects have no reasons at all to have common technical roots.

Equation (3) is true only when we consider the true standard deviations of the whole populations. In practice, these σ true values are estimated as s values from a limited number of results, what implies to deal with the number of degrees of freedom that are different from N_r and N_s. The use of the ANOVA (analysis of variances) methods is then needed to perform properly the computations.

As a conclusion of all of the upper, the test results that a given lab sends to the PT provider are not only governed by their bias, but also by which combination of equipment – operator – testing conditions that were used to perform the tests for PT, by the repeatability of tests and by chance with regard to inhomogeneities of samples.

In any cases, σ_{pt} is then always greater than σ_{BL}, what leads to α-risk lower than the expected 0.3 %, but also and consequently to increased β-risk.

In some cases, for example when $\sigma_r \gg \sigma_{BL}$ and only one test result is sent by each lab, the PT can become completely inefficient (see 3.4 here after).

In practice, in most cases:

1. σ_{iL} cannot be computed because each lab is requested to provide results obtained by only one operator, one test equipment set, performed in a short period of time (i.e. in repeatability conditions). Consequently, when each lab provides several test results, their computed standard deviation is s_r and does not include any contribution of σ_{iL};
2. Labs are requested to perform a few tests on a same sample or one test on each of a few distributed samples. In these conditions, σ_r and σ_H cannot be computed separately.

Consequently, in most cases only two standard deviations are governing the assessment:

1. An interlaboratory standard deviation that we call σ_L in our study, and that includes σ_{BL}, σ_{iL} and, when only one sample is provided, σ_H;
2. A repetition standard deviation that we call σ_{rPT} in our study, and that includes σ_r and σ_H when several samples are provided and one test per sample is performed.

This σ_{rPT}, can be determined from Equation (3), with respect to the design of the PT as defined by the PT provider (how many samples per participant, how many test results per sample, etc).

When only one test result from only one sample is provided per each participant (what in fact happens quite often), σ_{pt} is then the reproducibility standard deviation σ_R.

Note that we see here that PT providers could strongly improve their scheme and use ANOVA to separate all these standard deviations, but this goes far beyond the scope of this study and is not dealt with in this article.

In our study, we computed α and β-risks for σ_{rPT}/σ_L from 0.1 to 3 (corresponding to σ_{rPT}/σ_R from 0.1 to 0.95 that encompass the ratios actually encountered in practice) and for N_r (number of test results per lab) from 1 to up to 48. This latter number of repetitions is obviously quite too high to be encountered in practice. However, including it in our scheme made possible to investigate whether there could be of some benefit in some cases.

3. RESULTS AND DISCUSSIONS

3.1. General

All data of results is presented in figures. Detailed results are available in [13] (see particularly the annexes).

3.2. Pertinence of a ratio relating repeatability, interlaboratory standard deviation and number of test results per participant

To deal with the issue exposed at chapter 2.6, we defined a parameter λ as follows:

$$\lambda = \frac{\sigma_{rPT}}{\sigma_{iL} \times \sqrt{N_r}} \tag{4}$$

where σ_{rPT} is the standard deviation of sets of results of participants,

σ_{iL} is interlaboratory standard deviation,

and N_r is the number of test results per lab.

This parameter reflects the idea that the test results of each participant follow a Gaussian law which mean value is the bias and which standard deviation is $\sigma_{rPT}/\sqrt{N_r}$.

We found out that this parameter is valid to describe the full effect described in chapter 2.6, see Figure 1.

Figure 1 clearly shows that, for each number of participants, the σ_{rPT}/σ_{iL} curves are in extension of each other, so that a merge of these curves make sense, as shown in Figure 2.

3.3. Impact of the use of robust statistics

As seen in 2.5, [2] and [3] recommend not to use robust statistics when the number of participants is low because of lower efficiency while [10] and [11] did not confirm that this recommendation is useful.

Our computations confirmed that:

1. The α-risk is slightly increased when using robust statistics, what is consistent with the expected loss of efficiency in the determination of assigned values;
2. The β-risk is significantly reduced when using robust statistics, what is consistent with the better robustness of the assigned values.

In details, three cases were considered:

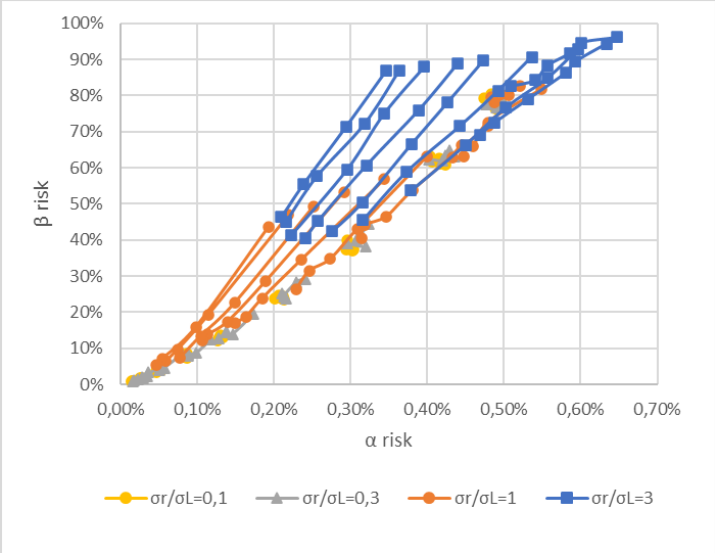

Figure 1. α- and β-risks for participants without outlier in function of λ and number of participants (from left to right, p = 5 - 6 - 8 - 10 - 13 - 16 - 20 - 25 - 30).

1. Comparison of risks for participants when no outlier is artificially introduced. In that case, occurrences of α-risk only relate to participants which are located at the tails of the Gaussian distribution, that is to say which bias is high by chance, without any technical reason for that;
2. Comparison of risks for not outlying participants when one outlier with a fixed bias is introduced in the population of participants. In that case, occurrences of α-risk also relate to participants which are located at the tails of the Gaussian distribution of the bias of participants;
3. Comparison of β-risks for an artificially introduced outlier which bias corresponds to a known z-score. By definition, the α-risk does not exist in that case (there is no risk to declare it as outlier while it is not).

Figure 3 shows the results of comparisons of risks when no introduced outlier is present. We observed that α-risk slightly increases while β-risk slightly decreases. However, both evolutions are not significant compared to the impact of the other factors (λ and N_p).

Figure 4 shows the results of comparisons of risks for main participants when one introduced outlier is present. We observed that α-risk slightly increases while the β-risk significant decreases when the λ factor is adverse (i.e. when $\lambda > 1$). In particular, we can see that even with 30 participants and $\lambda > 1$, not robust statistics completely fail to detect participants with $z > 3$ (β-risk > 90 %).

Figure 5 shows the results of comparisons of risks for an outlier. We observed that AlgoA is significantly more efficient to detect outliers even when PT conditions are adverse (i.e. when $\lambda > 1$ or when $N_p < 13$).

3.4. Impact of λ ratio

Figure 2 clearly shows that both α and β-risks decrease with λ until a certain value of λ that we evaluated to be 0.17, whatever the number of participants. When the critical value $\lambda = 0.17$ is reached, no further improvement of both α-risk and β-risk occur, whatever the number of repetitions. This can be clearly observed in Figure 1, where all results for $\sigma_{rPT}/\sigma_{iL} = 0.1$ are grouped in clusters (in orange on the figure).

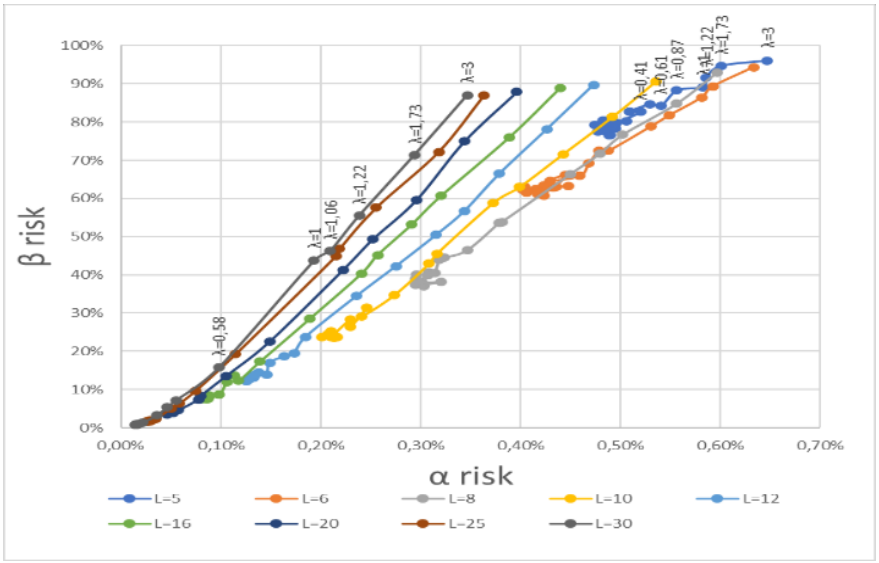

Figure 2. α- and β-risks for participants without outlier in function of the number of participants (L is the number of participants).

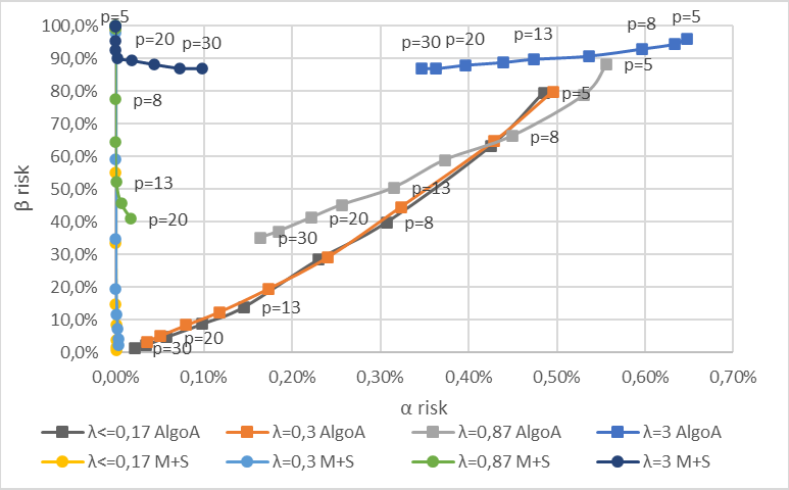

Figure 3. Comparison of α- and β-risks obtained with Algorithm A and with not robust statistics (*m* + *s*), for participants without any outlier in function of *λ* (*p* is the number of participants).

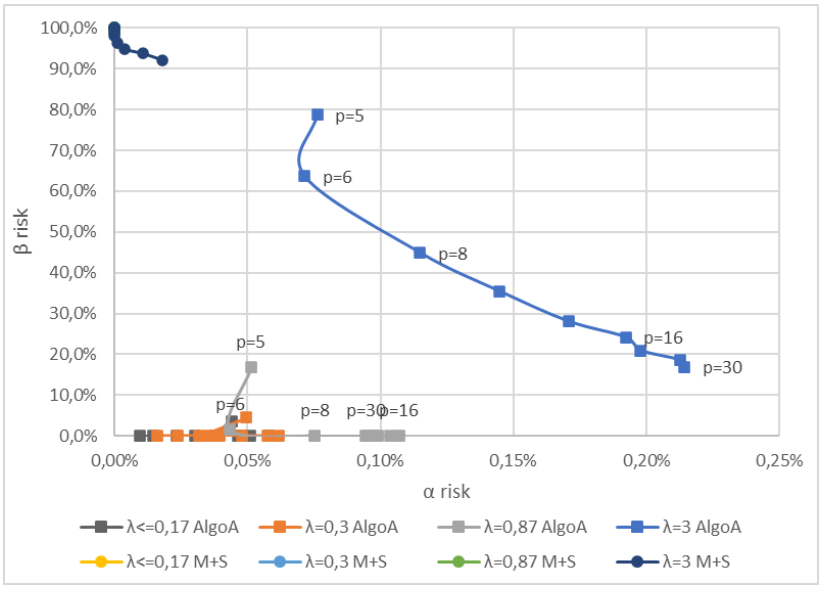

Figure 4. Comparison of α- and β-risks obtained with Algorithm A and with not robust statistics (*m* and *s*), for main participants when an outlier with *z* = 10 is present, in function of *λ* (*p* is the number of participants).

This also occurs for other cases (i.e. when an outlier is present) as shown in Table 3.

PT providers control neither σ_r nor σ_{iL}. These standard deviations only depend on the test method. But the PT providers do control N_r (the number of test results per lab) and hence do control λ (increasing N_r decreases λ, see (4)). They should use their historical data or literature to determine, σ_{rPT}/σ_R for each test method proposed for PT and they should use Table 3 to determine the minimum N_r values to optimize the PT programs. However, practical reasons may limit N_r (costs or impossibilities to produce or to transport the samples, costs or impossibilities for laboratories to perform a large number of tests).

As a conclusion, when N_r is chosen equal or superior to the value of Table 4, the best α and β-risks can be reached, according to the number of participants.

Further experiments are requested to understand the undergrounds of this $\lambda = 0.17$ constant. In particular, its variations in accordance with the definitions of H_0 and H_1 should be studied (see 2.1).

3.5. Discussion about α-risks

The theoretical α-risk with our definition of H_0 is 0.0027 / 0.95 = 0.28 % (probability that $|z_{calc}| > 3$ while $|z_{true}| < 2$). This risk is reduced by the impact of the repeatability, especially when the λ value is high (see 2.6). When the PT conditions are bad (i.e. $\lambda > 1$ or $N_p < 13$) the use of robust algorithms tends to increase α-risk while the use of mean value and standard deviation tend to decrease α-risk.

On the other hand, the comparison of Figure 3 and Figure 4 shows that the presence of outliers tends to decrease α-risk. Indeed, in those cases the σ_{pt} standard deviation is strongly over estimated, what significantly decreases the z-scores of all participants, including those of the opposite side of the distribution of results for which this effect is softened by the offset of the assigned central value.

In any cases, even in very bad PT conditions (i.e. $\lambda = 3$ and/or $N_p = 5$) α-risk always remains very low (less than 0.7 %), see Figure 3.

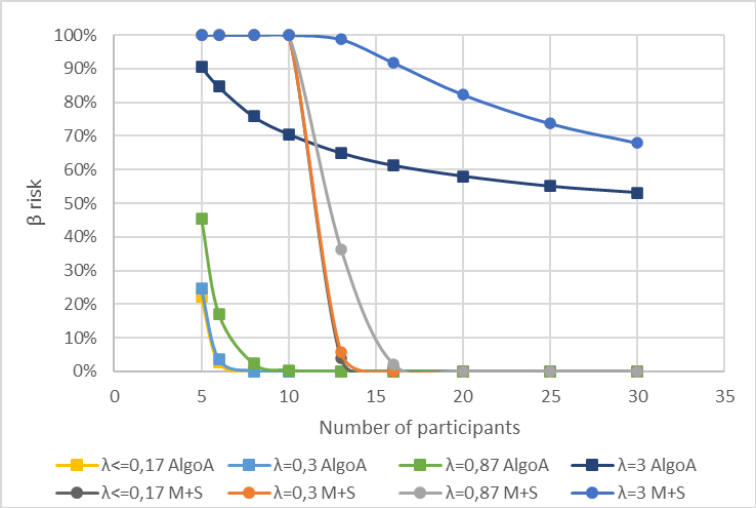

Figure 5. β-risks obtained with Algorithm A and with not robust statistics (m+s), for an outlier with z=10, in function of λ

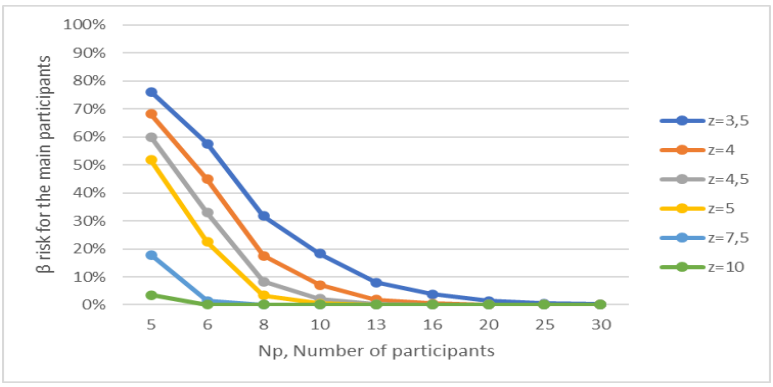

Figure 6. β-risks obtained with Algorithm A and λ=0.17 for the main participants when an outlier is present, in function of the outlier's z-score

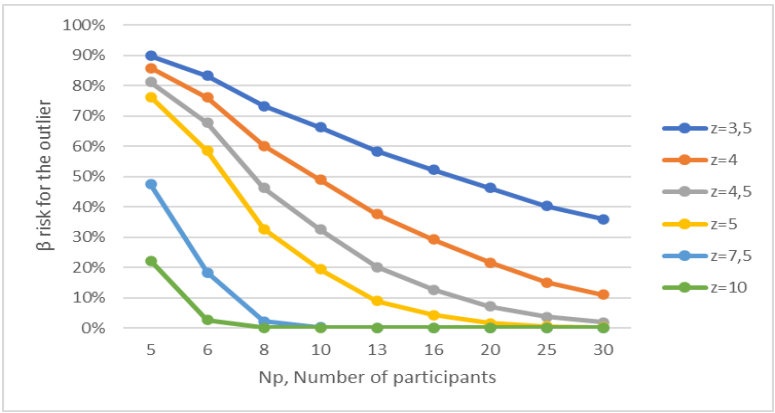

Figure 7. β-risks obtained with Algorithm A and λ=0.17 for an outlier in function of its z-score.

3.6. Discussion about β-risks

Whatever the situation (with or without presence of an outlier), β-risk is mainly governed by:

1. the λ ratio
2. and the number of participants.

Without any outlier, using $\lambda \leq 0.3$ and $N_p \geq 13$ is needed to get a β-risk less than 20 %, see Figure 3.

When an outlier whose $z = 10$ is present:

1. The β-risk for the main population is very close to 0 in almost all cases for which $\lambda \leq 0.9$, whatever N_p, see Figure 4;
2. The β-risk for the outlier is under control as soon as $\lambda \leq 0.3$ whatever the number of participants, see Figure 5

Figure 6 and Figure 7 show the β-risks respectively for the main participants and to the outlier in function of the outlier's z-score.

It is reminded that 0.3 % of the participants of the main population get z-scores $z < -3$ or $z > +3$. However, the H_0 hypothesis considers them as outliers, so that the H_1 hypothesis can be checked, i.e. a β-risk can be computed.

These figures show that 6 participants are enough to detect a strongly outlying participant (whose z-score is 10), while 30 participants are not enough to detect a slightly outlying

Table 3. Lower λ limits under which α and β-risks decrease anymore according to the number of participants (α and β in %, computed with Algo A).

		No outlier		Main participant when one outlier is present		Outlier	
N_p	λ	α (%)	β (%)	α (%)	β (%)	α (%)	β (%)
5	0.17	0.5	80	0.55	90	-	22
6	0.17	0.45	65	0.53	80	-	2
8	0.17	0.30	40	0.44	65	-	0
10	0.17	0.2	23	0.38	59	-	0
13	0.17	0.12	12	0.32	50	-	0
16	0.17	0.10	10	0.25	45	-	0
20	0.17	0.05	5	0.22	40	-	0
25	0.17	0.03	3	0.18	38	-	0
30	0.17	0.01	1	0.16	34	-	0

Table 4. Optimal number of repetitions for PTs, according to the σ_r/σ_{iL} and σ_r/σ_R ratios.

σ_r/σ_{iL}	σ_r/σ_R	N_r
≤ 0.17	≤ 0.17	1
0.3	0.29	3
0.42	0.39	6
0.59	0.51	12
1	0.71	35
3	0.95	310

participant (whose z-score is 3.5) even if PT conditions are optimal ($\lambda = 0.17$ and $N_p = 30$).

4. CONCLUSIONS

This study demonstrates that:

1. The ratio $\lambda = \sigma_r/(\sigma_{iL} \times \sqrt{N_r})$ is of main importance to control the efficiency of a PT scheme, even more than the number of participants. The PT providers should then care N_r, number of test results per participant that they request;
2. Even in adverse conditions, the α-risk is always very low (less than 0.7 %);
3. Robust algorithms improve the efficiency of the PT program (i.e. β-risk) at a slight expense on α-risk (which always remain very low). This comes from a significantly better estimation of the standard deviation of reference when an outlier is present among the participants;
4. A number of 6 participants is large enough to detect a strongly outlying participant provided that good PT conditions (i.e. low value of λ) are present;
5. PT with a low number of participants is (almost) always better than no PT at all.

Reference standards [2] and [3] recommend not to organise an ILC with less than 12 participants. This makes sense for [2], which goal is to determine the performance of a test method. It makes less sense for [3], which goal is to check the performance of a lab. Obviously, when no PT is organised, β-risk is 100%: any lab having a problem can never at all realise it! Consequently, for test methods that are performed by a little number of labs, it is obviously better to organise PT with 6 participants than nothing. In those cases, the PT provider should specially care the N_r it requests, to ensure a proper λ value and consequently assure an efficiency as good as possible.

ACKNOWLEDGEMENT

We acknowledge the institute CompaLab, PT organiser (www.compalab.org), for its scientific and technical support for this work.

REFERENCES

[1] ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories

[2] ISO 5725-2:2019 Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

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

[4] ISO 17043:2010 General requirements for proficiency testing

[5] ISO 3951-1:1999 Sampling procedures for inspection by variables — Part 1: Specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL

[6] David Luengo, Luca Martino, Mónica Bugallo, Víctor Elvira and Simo Särkkä, A survey of Monte Carlo methods for parameter estimation, EURASIP Journal on Advances in Signal Processing, Article 25, May 2020.
DOI: 10.1186/s13634-020-00675-6

[7] ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions

[8] ISO 16269-4:2010 Statistical interpretation of data — Part 4: Detection and treatment of outliers

[9] Maria Belli, Stephen, L. R. Ellison, Ales Fajgelj, Ilya Kuselman, Umberto Sansone, Wolfhard Wegscheider, Implementation of proficiency testing schemes for a limited number of participants, Accreditation and Quality Assurance 12 (2007), pp. 391–398.
DOI: 10.1007/s00769-006-0247-0

[10] Isao Kojima, Kakutoshi Kakita, Comparative study of robustness of statistical methods for laboratory proficiency testing, Analytical sciences, The journal of the Japanese Society for Analytical Chemistry 30 (2014).
DOI: 10.2116/analsci.30.1165

[11] Dimitris Tasmatsoulis, Comparing the Robustness of Statistical Estimators of Proficiency Testing Schemes for a Limited Number of Participants, Computation 10(3) (2022).
DOI: 10.3390/computation10030044

[12] ISO 5725-3:1994 Accuracy (trueness and precision) of measurement methods and results — Part 3: Intermediate measures of the precision of a standard measurement method

[13] Louis-Jean Hollebecq, β-risk in proficiency testing in relation to the number of participants, CompaLab, 2022. Online [Accessed 23 June 2023]
https://www.compalab.org/medias/files/publication-interne-risque-beta-en.pdf