17th International Congress of Metrology, 02003 (2015)
DOI: 10.1051/metrology/201502003

Comparaison de différentes stratégies d'évaluation de la performance dans les essais d'aptitude par comparaisons interlaboratoires.

Michèle Désenfant[1], Alexandre Allard[1], Béatrice Lalère[1], Sophie Lardy-Fontan[1], Véronique Le Diouron[1], Bénédicte Lepot[2], Eric Ziegler[3]

[1] *LNE, Laboratoire national de métrologie et d'essais, 1 rue Gaston Boissier, 75724 Paris cedex 15, FRANCE*
[2] *INERIS, Parc technologique ALATA, BP2 60550 Verneuil-en-Halatte, FRANCE*
[3] *Bipea, 189 rue d'Aubervilliers, 75018 Paris, FRANCE*

Résumé : Proficiency Testing (PT) uses interlaboratory comparisons to evaluate the performance of participants for tests or measurements. The statistical design and analytical techniques applied must be appropriate for the stated purpose of each PT. Most PT schemes compare the participant's deviation from an assigned value with a numerical criterion. Different scoring strategies are available and commonly used, as described in the International Standard ISO 13528, with a new version scheduled for publication at the end of 2015. The different strategies used to determine the assigned value and the criterion to calculate performance statistics are critical because they will lead to different interpretations. It is very important to consider whether the assigned value and its standard uncertainty as well as the criterion are independent of participant results, or are derived from the submitted results. The two strategies will be presented, compared and discussed in perspectives of the concept of comparability and the concept of compatibility that are defined in the International Vocabulary of Metrology (VIM). To do so, a certified reference material was deployed in two proficiency testing schemes. They aimed at determining a set of relevant pesticides- belonging to two chemical families: triazines and phenylurea - that are part of the list of priority substances of the European Water Framework Directive. Various methods to determine the assigned value from the results of participants were used : the algorithm A on one hand and the Q method and Hampel estimates on the other hand, as introduced in the new version of ISO 13528. Moreover, on the basis of the available estimators, the complementarity use of E_n and *Z score* in the perspectives of evaluating the performance of laboratories was addressed.

1 Objectif des essais d'aptitude (Proficiency Testing)

Les comparaisons interlaboratoires (CIL) sont définies comme l'organisation, l'exécution et l'évaluation d'essais ou de mesurages sur des échantillons identiques ou semblables par au moins deux laboratoires différents dans des conditions prédéterminées.

La mise en œuvre d'une comparaison interlaboratoires peut répondre à différents objectifs dont les trois principaux sont :
- Attribuer une valeur consensuelle à une caractéristique d'un objet (par exemple un matériau de référence) ;
- Estimer l'exactitude (justesse et fidélité) d'une méthode de mesure ;
- Evaluer l'aptitude des participants.

Un essai d'aptitude consiste à utiliser les comparaisons interlaboratoires pour évaluer la performance d'un laboratoire en matière d'essais ou de mesurages (norme 17043 [1]).

2 Les étapes d'analyse statistique d'un essai d'aptitude

Ce paragraphe s'appuie sur la norme *Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires* version actuelle et version en révision à l'ISO/TC 69 NF ISO 13528 : 2005 [2] et FDIS 13528 : 2015 [3].

La première phase consiste à s'assurer de la faisabilité de l'évaluation de la performance des participants, elle est bien sûr primordiale.
Elle comprend :
- l'étude d'homogénéité des entités soumises à essai
- l'étude de stabilité qui doit être contrôlée et garantie pendant toute la durée de la comparaison.
- l'analyse préliminaire des résultats des participants : retrait de résultats « très aberrant » (blunder), étude visuelle des résultats pour en vérifier la distribution et éventuellement mise en œuvre de tests de valeurs aberrantes.

La seconde phase consiste à déterminer la valeur assignée (*valeur attribuée à une grandeur particulière et reconnue, parfois par convention, comme ayant une incertitude appropriée à un usage donné,* définition issue de la norme NF ISO 13528) et son incertitude. Principalement 5 façons de déterminer cette valeur sont décrites dans la norme: par formulation, par matériau de référence, par un laboratoire, par valeur consensuelle de laboratoires experts, par valeur consensuelle des résultats des participants. Cette partie sera reprise pour développer l'importance de l'origine de cette valeur assignée dans la suite des calculs.

Enfin la troisième phase consiste à calculer la performance des participants. Différentes statistiques de performance sont possibles, la norme NF ISO 13528 en présentent cinq : l'écart noté *D*, le *Z* score, le *Z'* score, *Zeta* score, E_n score, impliquant ou non d'évaluer l'écart-type d'aptitude (noté σ_{pt}), prenant en compte ou non l'incertitude des résultats de chaque participant. Le Tableau 1 présente une synthèse des paramètres nécessaires à l'estimation de chacun des scores. Toutes les statistiques de performance ne sont pas évaluées à partir des mêmes paramètres.

Tableau 1. Tableau de synthèse des paramètres nécessaires aux calculs des scores de performance.

Paramètres / Performance.	x_i	$u(x_i)$	x_{pt} (Indép. / issue)	$u(x_{pt})$	σ_{pt}
D (ou D%)	●		●		
Z score	●		●		●
z' score	●		●	●	●
Zeta Score	●	●	●	●	
E_n	●	●	●	●	

x_i et $u(x_i)$: résultat et incertitude type du participant *i*

x_{pt} et $u(x_{pt})$: valeur assignée et son incertitude type,

σ_{pt} : écart-type pour l'évaluation d'aptitude

Les plus couramment utilisés sont, dans le domaine des essais d'aptitude, le *Z* score calculé à partir d'estimateurs robustes mis en œuvre sur les résultats des participants de la comparaison et, dans le domaine de la métrologie le score E_n, appelé également écart normalisé, estimé à partir du résultat et de l'incertitude du participant et d'une valeur externe de référence et son incertitude.

3 L'importance de la valeur assignée pour évaluer une performance

La valeur assignée d'un essai d'aptitude peut soit être une valeur indépendante des résultats des participants (une valeur de référence), soit une valeur estimée à partir des résultats des participants. En considérant que cette valeur assignée sert de base à la détermination des scores de performance des laboratoires et par conséquent à l'évaluation de leur aptitude, ceci est vraiment d'importance.

Dans le cas d'une valeur de référence externe pour valeur assignée, la performance d'un participant est élaborée indépendamment des résultats des autres participants, c'est à dire qu'on cherche à répondre à la question suivante : mon résultat est-il comparable à la valeur de référence ?

Dans le cas d'une valeur consensuelle des résultats de la comparaison pour valeur assignée, la performance d'un participant est élaborée en tenant compte de la dispersion statistique des résultats de tous les participants, on répond dans ce cas à la question suivante : mon résultat est-il compatible avec les résultats des autres participants ?

Dans ce dernier cas, l'analyse préliminaire des résultats de la comparaison est primordiale pour s'assurer de l'accord des participants avant de combiner leurs résultats dans une valeur consensuelle.

Cette distinction selon l'origine de la valeur assignée est schématisée sur la figure 1.

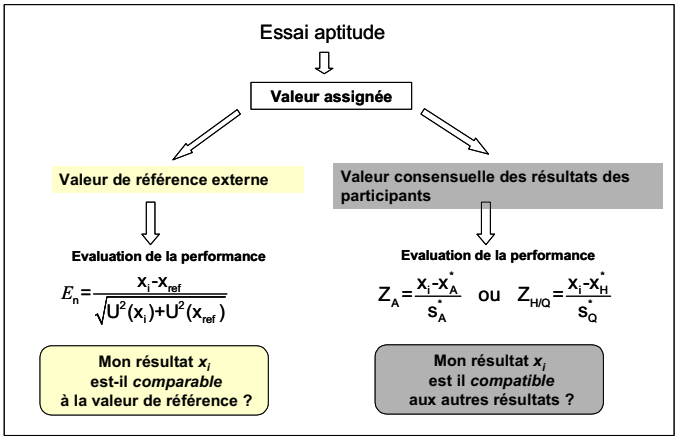

Figure 1 : Estimation de performance en fonction de l'origine de la valeur assignée

4 Valeur assignée estimée à partir des participants : les estimateurs robustes de la moyenne

4.1 Estimateur robuste classique : Algo A

Cet algorithme est présenté dans la norme NF ISO 13528 et mis en œuvre depuis de nombreuses années par les organisateurs de CIL.

Les calculs portent sur une démarche itérative ramenant les résultats éloignés aux bornes de limites calculées.

L'algorithme est le suivant :

1) Calculer les valeurs initiales de x^* and s^* selon:

x^* = médiane des x_i $\quad (i = 1, 2, ..., p)$

s^* = 1,483 médiane des $\left|x_i - x^*\right|$ avec $\quad (i = 1, 2, ..., p)$

2) Mettre à jour les valeurs de x^* et s^* comme suit

Calculer : $\delta = 1{,}5\ s^*$

Pour chaque x_i (i = 1, 2, ..., p), calculer:

$$x_i^* = \begin{cases} x^* - \delta & \text{quand } x_i < x^* - \delta \\ x^* + \delta & \text{quand } x_i > x^* + \delta \\ x_i & \text{sinon} \end{cases}$$

3) Calculer les nouvelles valeurs de x^* et s^* à partir de:

$$x^* = \sum_i \frac{x_i^*}{p}$$

$$s^* = 1.134 \cdot \sqrt{\sum (x_i^* - x^*)^2 / (p-1)}$$

On peut visualiser la fonction d'influence des résultats correspondant à cet algorithme A sur la figure 2.En fonction de leur éloignement à la moyenne, les résultats sont soit conservés, soit plafonnés.

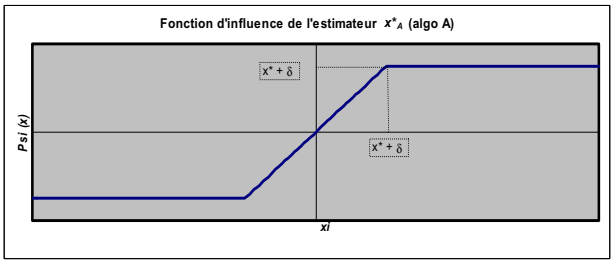

Figure 2 – Fonction d'influence des résultats pour l'estimateur robuste, Algo A

4.2 Estimateur robuste de Hampel (introduit dans la révision FDIS 13528)

Cet estimateur robuste est proposé comme alternative à celui de l'Algo A. Il est particulièrement apprécié pour son breakdown point, c'est à dire sa résistance aux points aberrants quand ils sont présents.
Il évite les problèmes de convergence de la procédure itérative de l'algo A qui itère simultanément le paramètre de position et dispersion. Se référer à « Robust estimates of location » pour plus d'information [4].

Calculer la moyenne robuste, x_H^*, en résolvant l'équation suivante :

$$\sum_{i=1}^{p} \psi\left(\frac{y_i - x_H^*}{s^*}\right) = 0$$

où

$$\psi(q) = \begin{cases} 0 & q \le -4{,}5 \\ -4.5 - q & -4.5 < q \le -3 \\ 1.5 & -3 < q \le -1{,}5 \\ q & -1.5 < q \le 1.5 \\ 1.5 & 1{,}5 < q \le 3 \\ 4.5 - q & 3 < q \le 4.5 \\ 0 & q > 4.5 \end{cases}$$

et s^* est l'écart-type robuste (selon la méthode Q présentée dans la norme mais non développée dans cet article centré sur la valeur assignée).

Si on regarde Figure 3, la fonction d'influence de ce nouvel estimateur, on voit cette fois-ci qu'en fonction de leur éloignement à la moyenne normalisée, les résultats sont soit conservés, soit plafonnés, soit pondérés faiblement, soit in fine pondérés à zéro.

(C.9)

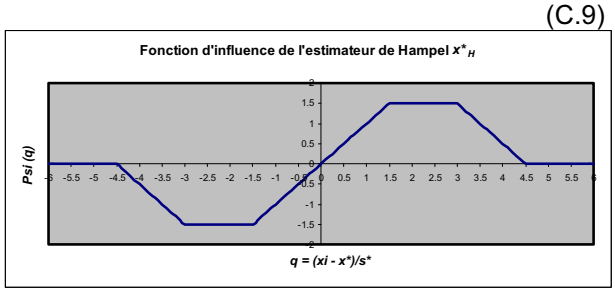

Figure 3 : Fonction d'influence des résultats pour l'estimateur robuste de Hampel

5 Illustration sur des données Eaux

5.1 Contexte et organisation

En Europe, la Directive Cadre Eau (DCE-2000/60/EC) et la Directive 2009/90/CE (Directive QA/QC) ont été mises en œuvre afin d'améliorer et de protéger la qualité de l'eau (eaux de surface, eaux côtières et eaux souterraines) mais aussi pour prévenir une nouvelle détérioration avant 2015. Dans ce contexte, l'évaluation de la qualité chimique des eaux repose sur la capacité des organismes en charge de la surveillance à assurer la comparabilité des données: comparabilité des données générées au sein d'un même laboratoire, comparabilité des données générées par les différents laboratoires mandatés dans les pays membres de l'union européenne, comparabilité des données générées dans le temps et dans l'espace.

En France, les laboratoires qui opèrent pour la surveillance des milieux aquatiques doivent être accrédités (selon le référentiel NF EN ISO/CEI 17025 : 2005 ou équivalent) et agréés par le MEDDE[1] sur la base du respect de critères de performances (limite de quantification, démonstration de l'aptitude en participant à au moins deux essais d'aptitude (lorsqu'ils existent) par an et obtention de $|Z$ score$|<2$.
Dans ce contexte, l'évaluation des performances de la méthode et sa validation sont des composantes essentielles des mesures que le laboratoire doit mettre en œuvre pour lui permettre de produire des données analytiques fiables. Ainsi, de nombreuses CILs sont régulièrement organisées en France. Les valeurs assignées, sont en général déterminées en prenant la

[1] Ministère de l'Ecologie, du Développement durable et de l'Energie.

moyenne (robuste) des résultats. Cependant, l'exactitude de ces valeurs assignées, n'est pas toujours établie, pouvant de ce fait entraîner des problèmes d'interprétation et d'exploitation des résultats par la présence de biais éventuels. L'utilisation d'une méthode primaire pour assigner la valeur de référence permet d'assurer la comparabilité des mesures.

En 2013, le LNE au travers du programme Aquaref appliqué à la surveillance des milieux aquatiques, s'est s'associé à deux organisateurs de comparaisons interlaboratoires le BIPEA et l'INERIS pour distribuer, en plus des échantillons de la campagne, une solution de référence certifiée en amont. Ces exercices portaient sur la détermination de pesticides (triazines et phénylurés) appartenant à la liste des substances prioritaires de l'état chimique (Directive, 2013/39/UE).

Un matériau de référence (certifié) MR(C) doit répondre à un certain nombre de caractéristiques : homogénéité, stabilité dans le temps, valeur fournie avec son incertitude.

Les valeurs de la solution de référence distribuée ont été déterminées après une étude d'homogénéité (n=20) et une étude de stabilité (suivi mensuel au cours de la première année). L'assignation de la valeur de référence a été établie au travers d'une comparaison inter laboratoires dits experts. Ainsi, la valeur assignée au MR(C) est la moyenne des résultats des laboratoires et l'incertitude-type est donnée par l'écart-type de reproductibilité.

Cette solution de référence a été distribuée au travers de 2 essais d'aptitude : le premier organisé par le BIPEA en juin 2013, le second par l'INERIS en décembre 2013. Le nombre de participants ayant accepté de quantifier cette solution de référence étant modeste, 11 laboratoires de la CIL BIPEA et 19 laboratoires de la CIL INERIS, les résultats des 2 campagnes ont été regroupés après avoir vérifié qu'il n'y avait pas d'écart dans l'organisation ni d'écart significativement différent de zéro dans les résultats.

Il est remarquable de noter que 29 laboratoires sur les 30 participants sont des laboratoires accrédités et 25 sont agréés par le MEDDE pour la surveillance DCE.

Seuls les résultats de la simazine seront présentés dans cet article.

5.2 Analyse des résultats

Le tableau 2 permet de synthétiser les éléments présentés ci-dessus en mettant en avant les différences dans l'évaluation de la performance selon l'estimateur considéré : *Z* score Algo A, *Z* score Q/Hample et nombre E_n. Les observations suivantes peuvent être faites:

- Pour la grande majorité des laboratoires, il y a une bonne cohérence entre les différents estimateurs de la performance;
- Le laboratoire 26 pour la simazine présente un *Z score*>2 avec une approche robuste basée sur l'Algorithme A alors qu'il présente un *Z score* = 2 avec une approche robuste basée sur Q/Hampel.
- Le laboratoire 2 présente un $E_n > 1$ alors que leurs |*Z* scores| sont < 2. Une explication possible semble être une sous estimation des incertitudes de mesure.
- A l'inverse, le laboratoire 18 présente des *Z scores* > 2 car sa valeur est plutôt éloignée des autres résultats, mais son incertitude associée est très grande donc il n'est pas détecté différent de la valeur de référence.

Code Labo	*En*	*Z score* Algo A	*Z score* Algo Q/H
1	-0.11	-0.67	-0.60
2	**1.14**	1.75	1.56
3	-0.27	-0.94	-0.84
4	0.23	0.22	0.19
5	-0.02	-0.39	-0.35
6	0.23	0.10	0.09
7	-	-0.80	-0.72
8	-0.37	-1.02	-0.91
9	-0.18	-0.88	-0.79
10	0.47	0.53	0.47
11	0.24	0.16	0.14
12	0.30	0.43	0.38
13	-	**-3.58**	**-3.20**
14	0.52	0.96	0.86
15	0.09	-0.12	-0.11
16	-0.03	-0.44	-0.39
17	0.47	1.02	0.91
18	0.83	**2.38**	**2.13**
19	**1.68**	**4.36**	**3.90**
20	-0.32	-0.94	-0.84
21	0.17	0.03	0.03
22	0.06	-0.19	-0.17
23	0.82	1.44	1.29
24	-	0.26	0.24
25	-0.11	-0.71	-0.63
26	**1.01**	**2.24**	2.00
27	-	-0.23	-0.21
28	0.21	0.04	0.04
29	**-1.16**	**-2.51**	**-2.25**
30	-0.07	-0.48	-0.43

Tableau 2 : Différents scores de performance pour les résultats de la Simazine

Les résultats sont également présentés sous forme graphique (Figure 4), permettant une comparaison rapide des différentes méthodes.

Dans le contexte des programmes de surveillance DCE, il est indispensable de pouvoir s'assurer des capacités des laboratoires devant opérer dans le cadre des marchés d'analyse dans le temps. Cette reconnaissance se fait dans le cadre de l'agrément délivré par le MEDDE. Pour qu'un laboratoire soit agréé, il doit entre autres, participer à au moins deux essais d'aptitude (lorsqu'ils existent) et obtenir un |*Z* score|<2.

Ainsi, le Z score obtenu par un laboratoire participant à une CIL, considérant l'approche robuste recommandée par la norme 13528, dépend des performances des laboratoires participant simultanément à lui à cette CIL. En d'autres mots, cette évaluation est donc contextuelle.

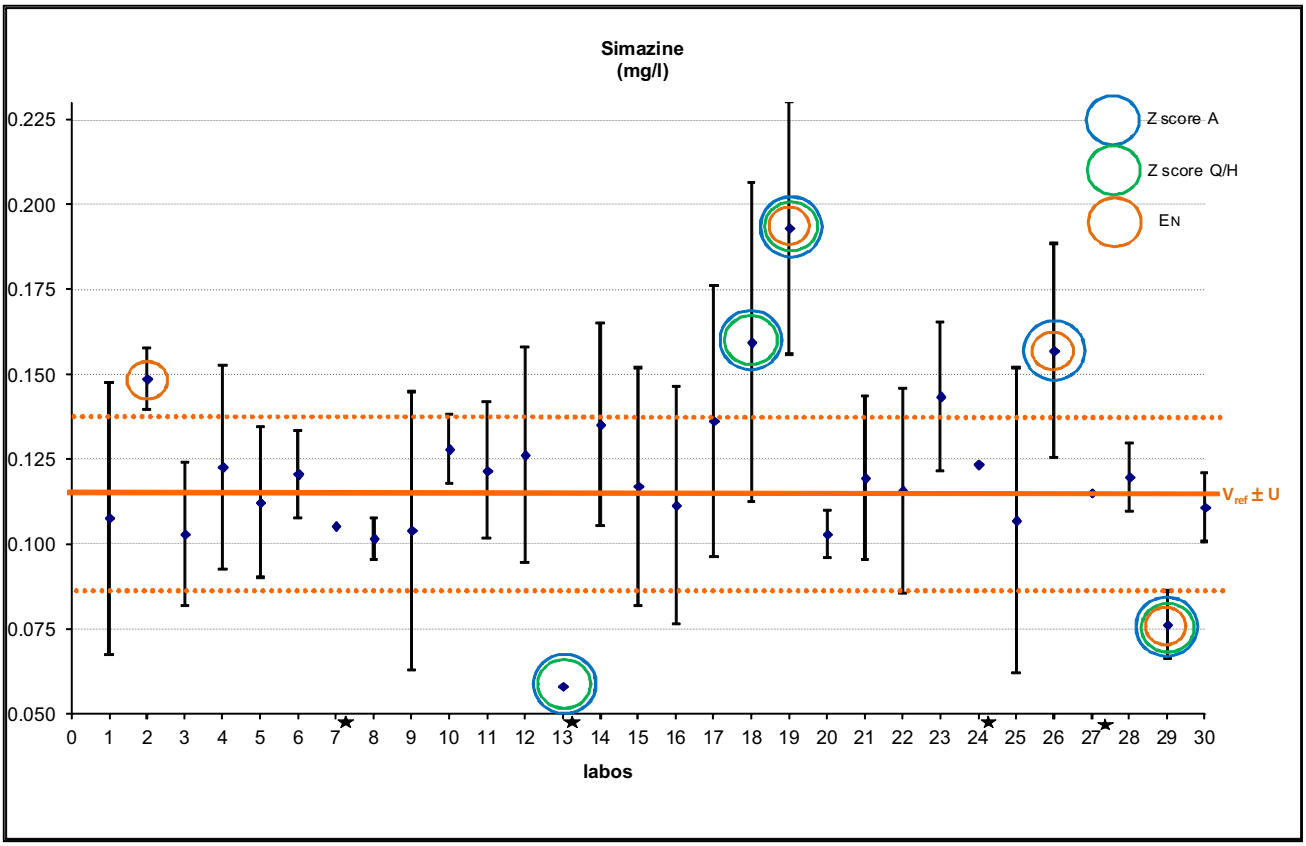

Figure 4 : Résultats de la simazine des 30 participants (valeur et incertitude élargie). Pour chaque laboratoire, mise en évidence des différences dans l'évaluation de la performance selon les estimateurs de valeur assignée : valeur de référence externe, valeur consensuelle Algo A ou Algo Q/H. (comme symbolisé par les cercles)

⋆ Indique que le laboratoire n'a pas fourni d'incertitudes et que par conséquent l'estimation *En* n'a pas pu être conduite. Le cercle symbolise que le laboratoire est en dehors des bornes d'acceptabilité : [-2 ; 2] pour le *Z* score et [-1 ;1] pour le nombre E_n

Pour illustrer également l'importance du choix de la statistique de performance, le Tableau 3 compare les performances selon que l'analyse porte seulement sur les 19 participants issus de la CIL INERIS ou sur l'ensemble des 2 circuits BIPEA et INERIS.
Dans les deux cas, les E_n sont bien sûrs les mêmes puisque ce score de performance est indépendant des résultats des autres participants. Par contre le *Z score* fait apparaître quelques différences, les participants 18 et 26 ont des scores supérieurs à 2 quand on prend tous les participants mais inférieurs à 2 quand on ne considère que le second circuit. Dans le cas du *Z score*, la performance dépend des résultats des autres laboratoires participants.

6 Conclusion

Ce travail expose l'importance de l'analyse des résultats (choix de la valeur assignée et du critère d'évaluation) dans le choix d'une campagne de comparaison interlaboratoires. Le plan statistique de traitement doit d'ailleurs toujours être annoncé par l'organisateur avant l'inscription des participants. Cet exercice démontre également l'intérêt du recours aux valeurs externes de référence comme valeur assignée pour démontrer l'évaluation de la justesse et l'aptitude des participants. On évalue la comparabilité des résultats qui apporte une information supplémentaire car indépendante des valeurs des autres participants.
Avec une valeur consensuelle des participants comme valeur assignée, on évalue la compatibilité avec les autres participants, c'est une autre information qui est utile et nécessaire pour estimer, par exemple, la dispersion et les statistiques de performance « d'une profession ».

Références

[1] NF ISO 17043 (2010) Evaluation de la conformité – Exigences générales concernant les essais d'aptitude

[2] NF ISO 13528 (2005) Méthodes statistiques utilisées dans les essaie d'aptitude par comparaison interlaboratoires
[3] ISO/FDIS 13528 (2015) Méthodes statistiques utilisées dans les essaie d'aptitude par comparaison interlaboratoires
[4] D.F. Andrews, P. J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers and J.W Tukey, Robust estimates of location, Princeton Legacy Library

	En	*En*	Z score Algo A	Z score Algo A
Code Labo	30 lab.	19 lab.	30 lab.	19 lab.
1	-0.11		-0.67	
2	**1.14**		1.75	
3	-0.27		-0.94	
4	0.23		0.22	
5	-0.02		-0.39	
6	0.23		0.10	
7	-		-0.80	
8	-0.37		-1.02	
9	-0.18		-0.88	
10	0.47		0.53	
11	0.24		0.16	
12	0.30	0.30	0.43	0.19
13	-	-	**-3.58**	**-3.03**
14	0.52	0.52	0.96	0.62
15	0.09	0.09	-0.12	-0.25
16	-0.03	-0.03	-0.44	-0.51
17	0.47	0.47	1.02	0.67
18	0.83	0.83	**2.38**	1.76
19	**1.68**	**1.68**	**4.36**	**3.36**
20	-0.32	-0.32	-0.94	-0.91
21	0.17	0.17	0.03	-0.13
22	0.06	0.06	-0.19	-0.31
23	0.82	0.82	1.44	1.01
24	-	-	0.26	0.06
25	-0.11	-0.11	-0.71	-0.72
26	**1.01**	**1.01**	**2.24**	1.65
27	-	-	-0.23	-0.34
28	0.21	0.21	0.04	-0.12
29	**-1.16**	**-1.16**	**-2.51**	**-2.18**
30	-0.07	-0.07	-0.48	-0.54

Tableau 3 : E_n et Z score calculés en considérant dans un cas 19 participants à la CIL et dans l'autre 30 participants