Accred Qual Assur (2006) 10: 461–465
DOI 10.1007/s00769-005-0021-8

D. Kisets

Performance indication improvement for a proficiency testing

Received: 25 October 2004
Accepted: 7 August 2005
Published online: 22 November 2005

Presented at 15th International Conference on Quality in Israel, November 2004

D. Kisets (✉)
The National Physical Laboratory of Israel (INPL),
Givat Ram,
91904 Jerusalem, Israel
e-mail: kisets@netvision.net.il
Tel.: +972-25664-976
Fax: +972-26520-797

Abstract The paper discusses peculiarities of Z scores and E_n numbers, which are most often used for the treatment of proficiency test data. The important conditions of proper usage of these performance indicators and their improvement are suggested on the basis of systematic approach, on the idea of accuracy classification, and on some principles of optimality borrowed from information theory. The author believes that this paper may be of interest and practical value for all those engaged in applied metrology, specifically in the field of developing of and participating in proficiency testing programs, and in the activity connected with accrediting testing and calibration laboratories.

Keywords Interlaboratory comparisons · Z scores · E_n numbers · Informational optimality

Introduction

Proficiency testing being carried out to assure the quality of test and calibration and demonstrate a laboratory competence is based on certain quantitative criteria of judging the quality of results obtained in the process of interlaboratory comparisons that have been called performance indicators. Among various statistics [1, 2] Z scores and E_n numbers are most often used as that kind of indicators and are being expressed in symbols of [1] as follows:

$$Z = |x - X|/s; \tag{1}$$

$$E_n = |x - X|/\left(U_{\text{lab}}^2 + U_{\text{ref}}^2\right)^{1/2}, \tag{2}$$

where x and X are the result of participating laboratory and the assigned value respectively (for E_n numbers X is usually the result obtained by reference laboratory); s is the estimate or measure of variability (standard deviation as a rule); U_{lab} and U_{ref} are the expanded uncertainty of the result of participating and reference laboratory respectively. The usage of performance indicators is specified [1] as $Z{\leq}2$ = satisfactory, $2{<}Z{<}3$ = questionable, $Z{\geq}3$ = unsatisfactory; and $E_n{\leq}1$ = satisfactory. The vagueness of these conditions (estimation criteria) for Z scores, as well as measurement uncertainties (if bearing in mind their absence in Z scores, and the way of interpreting E_n numbers) influence the quality of estimation results.

Whilst meeting with the recommendations of ISO/IEC Guide 43-1 [1] and currently developing ISO/DIS 13528 [2], a correct choice of proper performance indicator sometimes does also present a problem, which stems from the lack of well-founded criteria and methods of proving the choice. In measurement comparison schemes, for instance, the traditional use of E_n numbers for some combinations of measurement uncertainties of participating laboratories may lead to erroneous results, whereas the non-traditional transition to Z scores might in some cases improve the situation. Likely occasions of preferable usage of E_n numbers instead of Z scores also cannot be excluded from the practice.

Methodologically, either a performance indicator or (and) its applying is far from being perfect. In this connection the ignoring of measurement uncertainty (as pointed out in [3]), and the lack of certainty in estimation condition (non-optimal estimation) inherent in the usage of Z scores, the way proposed, for example, in [4] and aimed at allowing for the uncertainties for a case of applying in-house reference materials, is noteworthy that actually has led to using the performance indicator similar E_n numbers, rather than traditionally Z scores. In [2] the usage of two modernized performance indicators (so-called

z'-scores and zeta-scores): $z' = |x - X|/(s^2 + u_x^2)^{1/2}$ and $zeta = |x - X|/(u_x^2 + u_X^2)^{1/2}$, where u_x and u_X are standard uncertainties of the results of participating laboratory and the assigned value respectively, has yet been also stipulated for improving estimation capability.

As for E_n numbers, the consideration of this performance indicator in terms of statistics as being derived from Z scores [2] is not matching the metrological nature of the comparison of calibration laboratories with the reference one. The estimation reliability when using E_n numbers depends on both how the absolute error $|x{-}X|$ is normalized with respect to U_{lab} and U_{ref}, and of the correctness of allowing for these uncertainties. The most reliable normalization $E_{nr}{=}|x{-}X|/U_{\mathrm{lab}}$ is achievable when $U_{\mathrm{lab}} \gg U_{\mathrm{ref}}$. Irrespective of the last condition, E_{nr} has been used [5] through 1992 [6]. Expression (2), based on comparing the difference between the results obtained by laboratories and the uncertainty of calculating the difference as the criterion of proper estimating the competence of a laboratory, is incorrect in principle. A formal use of (2) may in some cases distort estimations and decrease their reliability, therefore the declared in [7] convenience of the method based on (2) is not a sufficient substantiation for its practical usage.

For all reasons given above, the present work is focused on the following two purposes: (1) the determination of practical conditions of correct applying Z scores and E_n numbers based on the uncertainties of participating laboratories, and (2) the improvement of performance indicators as such; this concerns the expression and estimation criteria for E_n numbers, and the optimal estimation criterion for Z scores. For achieving these goals the paper suggests classification approach. It does not discuss problems of designing and interpretation of proficiency tests, of determining the assigned value and its uncertainty, of the standard deviation for proficiency assessment and calculation of performance statistics that have been circumstantially presented in ISO/DIS 13528 [2]. In the author's opinion, the approach and methods proposed in the paper are not in contradiction with these problems, but in complement of one another.

Classification approach and general criterion of applying *Z* scores and *E*$_n$ numbers

The two performance indicators under consideration are based on different approaches. The difference $|x{-}X|$ for Z scores belongs to the system of statistical treatment of test or measurement result, whereas for E_n numbers it is considered as the estimated measurement error when comparing laboratories, one of which is the reference laboratory. For the best usage, the performance indicators should be considered as being intended for the different hierarchical levels of accuracy in principle. It is possible to tell that unlike Z scores E_n numbers is based on the principle of etalon (the less ratio $U_{\mathrm{ref}}/U_{\mathrm{lab}}$ the more reliable estimation). Thus the reliable use of E_n numbers demands the higher level of accuracy classification for a reference laboratory in comparison with other participants, whereas the Z scores normally (X and s are derived from participants' results) is applicable for the laboratories of the same accuracy level. This level represents the range of relative values farther derived and substantiated.

No matter how many laboratories participate in an intercomparison, the final judging for each participant is the result of comparing either directly with reference lab or with some conditional (virtual) reference lab, to which the assigned value together with its uncertainty could be attributed. Clearly, this is the process of classifying the laboratories by means of certain numerical value—the classification factor (as estimation criterion). The same approach is true when classifying measurement capabilities of laboratories, by analogy with the accuracy classification of measuring instruments.

In general view, the classification approach in applying the performance indicators requires collating the uncertainties U_a and U_b of any two (A and B) from the number of participating laboratories, one of which can specifically be the reference or conditional reference laboratory, for deciding if they belong to the same classification level or not, and on this basis to decide what performance indicator is particularly applicable. In order to do that, the relative influence of U_a and U_b (for the model of errors) or U_a^2 and U_b^2 (for the exclusively statistical model) on the estimation quality may be presented as weights $K_a{=}U_a/(U_a{+}U_b)$ and $K_b{=}U_b/(U_a{+}U_b)$, or $K_a = U_a^2/(U_a^2 + U_b^2)$ and $K_b = U_b^2/(U_a^2 + U_b^2)$ respectively, which are analogous to the estimates of probability forming the complete group of independent events ($K_a{+}K_b{=}1$). Such an analogy gives us the exclusive chance of applying the optimal selection model, borrowed from information theory [8], for classifying components of the group onto informative and redundant ones.

The ratio $\rho{=}\min(K_a/K_b)^{\pm 1}$ represents special coefficient, the best value of which is evaluated for the optimal (necessary and sufficient) rational or irrational positive number φ_o of components ($2{\geq}\varphi_o{\geq}1$) [9]. For the least certain situation about allowing or ignoring the lesser of two components (50% confidence), φ_o matches the following equation:

$$\varphi_o = \exp(-K_a \ln K_a - K_b \ln K_b) = 1.5 \qquad (3)$$

Accordingly, the optimum coefficient ρ_o of the weightiest component K (equals to either K_a or to K_b) as function of ρ is determined and approximately calculated as follows:

$$\begin{aligned}\rho_o &= \arg\left\{\exp\left[-\left(\frac{K}{K+\rho K}\right)\ln\left(\frac{K}{K+\rho K}\right)\right.\right.\\ &\qquad \left.\left. -\left(\frac{\rho K}{K+\rho K}\right)\ln\left(\frac{\rho K}{K+\rho K}\right)\right] = 1.5\right\}\\ &= \arg\left\{\exp\left[-\left(\frac{1}{1+\rho}\right)\ln\left(\frac{1}{1+\rho}\right)\right.\right.\\ &\qquad \left.\left. -\left(\frac{\rho}{1+\rho}\right)\ln\left(\frac{\rho}{1+\rho}\right)\right] = 1.5\right\} = 1/2\pi \qquad (4)\end{aligned}$$

The next expressions (5) and (6) result from the replacement of weights in formulas (3) and (4) by respective ratios of uncertainties.

For the model of errors (E_n numbers):

$$\rho_o = \min(U_a/U_b)_o^{\pm 1} = 1/2\pi, \tag{5}$$

For the statistical model (Z scores):

$$\rho_{os} = \rho_o^2 = \min\left(U_a^2/U_b^2\right)_o^{\pm 1} = 1/2\pi, \tag{6}$$

Expression (5) has been called *optimum accuracy coefficient* which, being the factor of relative classification, is also of fundamental significance for creating optimal systems of accuracy classification [10]. As the classification factor, the optimum accuracy coefficient is the first and most general criterion of judging, with the aid of which one can realize whether two compared laboratories belong to the same accuracy level (when $\min(U_a/U_b)_o^{\pm 1} \geq 1/2\pi$). If not, the use of E_n is the only correct decision. However, in case of laboratories of the same level of accuracy, but of different measurement capabilities, the situation is not so definite to make decision what performance indicator preferably to use; and the nearer $\min(U_a/U_b)_o^{\pm 1}$ to $1/2\pi$ the more it is indefinite. The more circumstantial estimation is achievable by using (5) and (6) expressions jointly; the details are given in the following section.

Peculiarities of applying *Z* scores and *E_n* numbers

Allowing for expressions (5) and (6), Fig. 1 reflects peculiarities of applying the performance indicators, namely the possibilities and preferences of their usage depending on decreasing accuracy coefficient $\rho = \min(U_a/U_b)^{\pm 1}$, i.e. the increase of quality of reference laboratory in carrying out intercomparisons.

In accordance with Fig. 1 there is a variety of possibilities in applying Z scores and E_n numbers depending on the accuracy coefficient regarding the laboratories undergoing comparison. Among the possibilities, special attention may be drawn to the following practical situations:

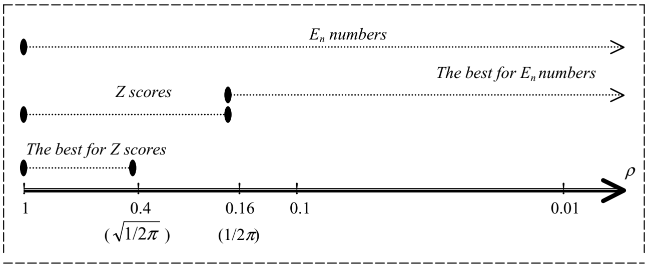

Fig. 1 Schematical illustration of boundaries and areas of possible and best usage of E_n numbers and Z scores when comparing two laboratories in the framework of proficiency testing programmes depending on the accuracy coefficient $\rho = \min(U_a/U_b)^{\pm 1}$

1. If all laboratories participating in the comparison meet the condition $U_{\text{ref}}/U_{\text{lab}} \geq 1/2\pi$ (where U_{ref} is somehow appointed, even from the participants), and their performance estimation determined by E_n numbers is unsatisfactory, then a final decision ought to be done using Z scores.
2. A laboratory with the uncertainty less than $1/2\pi$ part in the comparison with others ought to be appointed as the reference laboratory. In this case E_n numbers should be used only.
3. In case of two or more laboratories, uncertainties of which match the situation (1), for comparing their results it is reasonable to apply the Z scores model, whereas the model of E_n numbers—for the others. In this instance any of laboratories that matches the situation (1) might be chosen as the reference laboratory, and the less uncertainty of a laboratory, the more reliable choice will be made.

The significant limitation of applying E_n numbers in the area of Z scores (within the range $1/2\pi \leq \rho \leq 1$) is the often observed estimation uncertainty due to instability (drift) of the result of testing or calibration during the intercomparison period, even if the instability is within admissible limits of measurement uncertainty. Essential difference in the sensitivity of these two performance indicators to the instability of artifact is clearly demonstrated (Fig. 2) for the case of the same uncertainty (U) of two laboratories undergoing the comparison, and for the above-mentioned maximum permissible instability. In this case, the expressions $E_n = 0.707(\Delta + U)/U = 0.707(\Delta/U + 1)$ and $Z = [\Delta^2 + (U/2)^2]^{1/2}/(U/2)^2 = (4\Delta^2/U^2 + 1)^{1/2}$, where $\Delta = |x - X|$, were obtained using formulas (2) and (1) and the coverage factor 2 in determining U.

Optimum estimation criterion for *Z* scores

The model of Z scores has an affinity with well-known statistics used for detecting the outlying observations [11],

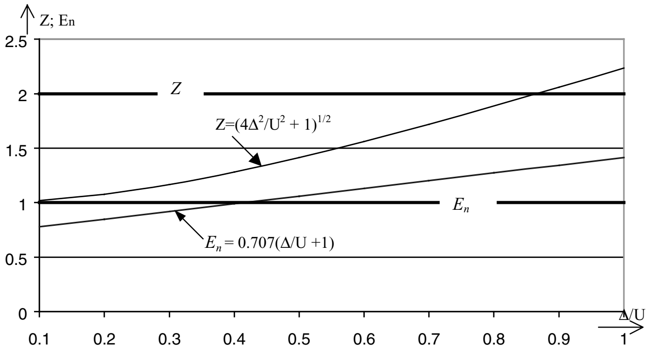

Fig. 2 Illustration of the advantage of Z scores regarding the influence of instability of measurand upon the results of intercomparison in the case of the same uncertainty (U) of two laboratories undergoing the comparison: the area of insensitivity of Z scores to the ratio Δ/U, where $\Delta = |x - X|$, for which the decision of satisfactory estimation ($Z \leq 2$) is true, is significantly greater than for E_n numbers in the respective area ($E_n \leq 1$)

where the quality of estimation depends upon the level of confidence chosen without taking into account the systematic character of quality estimation for system-defined measurements [12]. This method, despite relying on practical experience, is the clear-cut example of usage of exclusively statistical estimation towards the quality of estimation information that is a not always correct.

In reality the information obtained with Z scores should be optimum, and the aim of the optimization is to find out the best value (Z_o) of the ratio between $|x{-}X|$ and s as the optimal estimation criterion. However, until now it has been difficult to solve this problem. The difficulty can be successfully overcome by using the proposed above criterion of optimal classification for the statistical model as follows:

$$|x - X|^2/s^2 \leq 2\pi \;\text{ or, approximately, } |x - X|/s_c \leq 2.5 \qquad (7)$$

Applying this principle, the following convenient expression for comparing calibration laboratories through their uncertainties by means of Z scores may be recommended:

$$Z_o = |x - X|/[(U_{\text{lab}}/k_{\text{lab}})^2 + (U_{\text{ref}}/k_{\text{ref}})^2]^{1/2} \leq 2.5, \qquad (8)$$

where k_{lab} and k_{ref} = the coverage factors in determining the uncertainties. The left part of this condition is analogous as specified in [2] for zeta-scores, just expressed by using expanded uncertainties.

Thus, obtained condition (8) differs from presently used ones in principle since the new is characterized by the strictly established optimal estimation criterion (classification level = 2.5). However, it should be emphasized that this criterion, excluding vagueness in the interpretation of estimation results, is mostly useful and preferably applicable on the final stage of intercomparison. The presently used criteria may perform another function—the maintenance of the process on intermediate stages, if any, in order to consider participant's result in terms of (2) as giving an "action signal" (when Z score above 3.0 or below −3.0), or a "warning signal" (when Z score above 2.0 or below −2.0). Naturally, in this case being applied together the proposed criterion and the presently used ones do not contradict, but are the complement of one another.

Modernized expressions for E_n numbers

The improved expression for E_n numbers uses the accuracy coefficient $\rho{=}U_{\text{ref}}/U_{\text{lab}}$ as follows [13]:

$$E_{ni} = E_{nr} = |x - X|/U_{\text{lab}} \leq (1 - \rho); \qquad (9)$$

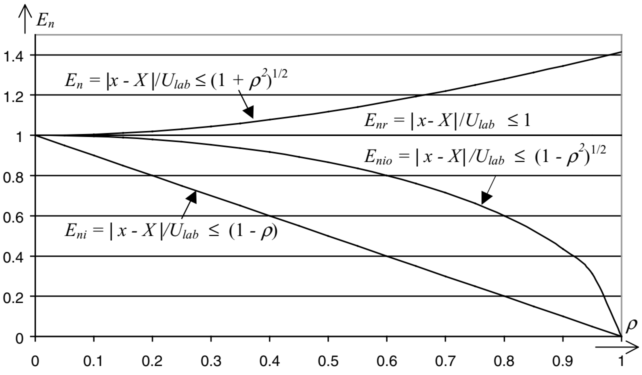

Fig. 3 Different expressions of E_n numbers as functions of accuracy coefficient $\rho{=}U_{\text{ref}}\,/U_{\text{lab}}$. Essential difference between E_n and E_{ni} (or E_{nio}) results from the opposite approaches to the estimation: unlike E_n, the increasing of accuracy coefficient leads to more strict acceptability condition for E_{ni} and E_{nio} as the compensation to lowering the quality of reference laboratory

This condition is based on the comparison of maximal errors by using the modulus of the uncertainties together with the discrepancy of measurement results. The maximum estimation reliability is achievable with such modular approach when compare the uncertainty U_{lab} with the maximal (due to U_{ref}) difference of measurement results, i.e. $\max|x{-}(X{\pm}U_{\text{ref}})|{=}|x{-}X|{+}U_{\text{ref}}$. The less reliable but the more optimistic estimation is when in formula (9) the criterion $\sqrt{1-\rho^2}$ is used instead of $(1{-}\rho)$.

The graphical illustration of all discussed expressions for E_n numbers (Fig. 3) demonstrates that, as distinct from commonly used E_n, the improved expressions E_{ni} and E_{nio} ensure toughening of the requirement to the difference $|x{-}X|$ due to the uncertainty of reference laboratory for achieving satisfactory result of the comparison.

For achieving the best estimation, the uncertainty U_{lab} in the expression (9) ought to be taken with the optimum level of confidence $C_o{=}(1{-}0.5\rho_o){\times}100\%{=}(1{-}1/4\pi){\times}100\%{=}92\%$. Because C_o and ρ_o characterize the absence of redundant estimation information, instead of (9) the following informatively more accurate expressions are preferable:

$$E_{ng} = E_{nr} = |x - X|/U_{\text{lab}} \leq (k_i/k_o)(1 - \rho); \qquad (10)$$

$$E_{ngo} - E_{nro} = |x - X|/U_{\text{lab}} \leq (k_i/k_o)\sqrt{1 - p^2}, \qquad (11)$$

where k_l = the coverage factor in determining the uncertainty U_{lab} (usually $k_l{=}2$); $k_o{=}1.75$ is the coverage factor, which matches C_o.

Example

Here is the example of determining the suitable model for comparing the seven testing or calibration laboratories, the results and uncertainties of which are presented (in conditional units) in the Table 1. The estimates of assigned value (X) and the variance (s) were calculated as follows: $X = (\sum_{i=1}^{7} X_i)/7 = 5.44$, and $s = [\sum_{i=1}^{7}(x_i - X)^2/6]^{1/2} = 0.15$.

Table 1 Data and results for the presented example of seven laboratories

Lab #	x_i	U_i	Z_i	$E_{nr} = \lvert x_i - x_2 \rvert / U_i$	$(1-\rho)$	$\sqrt{1-\rho^2}$
1	5.61	0.25	1.14	0.64	0.72	0.99
2	5.45	0.03	0.06	–	–	–
3	5.51	0.10	0.46	0.6	0.7	0.95
4	5.40	0.05	0.26	1	0.4	0.8
5	5.59	0.20	1.00	0.7	0.85	0.99
6	5.32	0.20	0.80	0.65	0.85	0.99
7	5.19	0.25	1.66	1.04	0.88	0.99

The results of Z_i calculation (Table 1) show all participants meet the condition (7). However, the uncertainty of Lab #2 is less than 1/2π of the uncertainty of any other participating laboratory; therefore this laboratory can be truly recognized as the reference laboratory regarding the rest of participants. Thus, this is a good case for further reasonable applying the E_n numbers instead of Z scores in order to ensure the correct estimation. Calculated by the expressions (10) and (11) E_n data (here k_l=k_2=2) demonstrate the non-conformance of the two (#4 and #7) laboratories by E_n numbers criterion, as well as the necessity of using just this criterion to prevent erroneous estimation conclusions. Thus, for this case the use of E_n numbers, instead of Z scores has enabled to evade a misconception regarding above-mentioned two laboratories. It should be recommended to decrease their best measurement capabilities by means of increasing the respective uncertainties.

Conclusions

1. Applying the performance indicators, the positive decision (in terms of acceptance) tells about the two in principle distinct estimation results, namely: (1) the belonging of compared laboratories to the same accuracy class—in case of Z scores model, and (2) the acceptable measurement error of a laboratory against the reference laboratory—in case of E_n numbers.
2. The combined application of improved Z scores model and the optimum conditions for the uncertainties of participating laboratories is proposed that enables recognizing the suitable model (Z scores or E_n numbers or even both of them) to use in a proficiency-testing situation. This makes it possible to achieve an optimum estimation when calculating and judging the quality of results in interlaboratory comparisons.
3. Proposed approach gives a chance to all those responsible for the fulfillment of interlaboratory comparisons to carry out the correct treatment of test, measurement and estimation data concerned. In the author's opinion, the current philosophy of development and operation of proficiency testing schemes [1] and methods being used [2] can be materially improved on the basis of the suggested ideas. First of all this concerns the simple and effective procedure for the optimal treatment of intercomparisons data that may be easily developed and unified on the basis of this paper as well. The classificational analysis of participating laboratories regarding their uncertainties represents an integral part of such a procedure.
4. The availability of different (proposed and presently used) philosophies in realizing estimation criterion of Z scores for the different stages of intercomparison process is the major inference drawn from their consideration. Acknowledging the practical significance of currently developing ISO/DIS 1358 [2], this may also pay attention to its improvement.

References

1. ISO/IEC Guide 43-1 (1997) Proficiency testing by interlaboratory comparisons—Part 1: Development and operation of proficiency testing schemes. ISO, Geneva
2. ISO/DIS 1358 (2002) (Provisional Version) Statistical methods for use in proficiency testing by interlaboratory comparisons, Committee identification: ISO/TC 69/SC 6. Secretariat: JISC/JSA
3. Kuselman I, Papadakis I, Wegscheider W (2001) Accred Qual Assur 6:78–79
4. Kuselman I, Pavlichenko M (2004) Accred Qual Assur 9:387–390
5. WECC Doc. 15 (1987) WECC International Measurement Audits
6. WECC Doc. 15 (1992) WECC Interlaboratory Comparisons
7. EAL-R7 (1995) EAL Interlaboratory Comparisons
8. Shannon CE, (1948) Bell Syst Tech J 27:379–423
9. Kisets D (2003) Information sufficiency of an uncertainty budget. Proceedings of 2nd International Conference on Metrology, November, Eilat, Israel.
10. Kisets D (1997) OIML Bull XXXVIII (2):30
11. ASTM E178-94 (1995) Standard practice for dealing with outlying observations, Annual book of ASTM Standards, Section 14, Vol. 14.02. pp 91–107
12. Kisets D (2000) Information cyclicity behind two major trends in metrology, Proceedings of The International Conference in Metrology, Jerusalem, pp 315–320
13. Kisets D (1997) OIML Bull XXXVIII(4):11–20