

Review of the new edition of ISO 13528

Siu Kay Wong¹

Received: 25 April 2016 / Accepted: 9 May 2016 / Published online: 2 June 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract The international standard ISO 13528 is an important reference document for proficiency testing providers and testing laboratories alike, which provides guidance on the use of statistical methods in proficiency testing. With a view to bringing the document in harmony with the ISO/IEC 17043:2010, the first edition of ISO 13528 was revised and the new edition was published in September 2015. This paper attempted to review and discuss the major changes introduced in the new edition of ISO 13528, particularly the part on statistical design with respect to the objective of proficiency testing scheme and the considerations on assigned value determination and performance scoring. Moreover, procedures provided in the new edition of ISO 13528 for qualitative proficiency testing schemes and various robust statistical methods would also be briefly discussed in the paper.

Keywords Proficiency testing scheme · Statistical design · Statistical methods

Introduction

The first edition of ISO 13528 “Statistical methods for use in proficiency testing by interlaboratory comparisons” was published in 2005 as a complement to the then ISO/IEC Guide 43-1:1997 “Proficiency testing by interlaboratory comparisons—Part 1: Development and operation of

proficiency testing schemes.” In 2010, ISO/IEC 17043 [1] was published to replace Parts I and II of ISO/IEC Guide 43:1997, which has preserved and updated the principles for the operation of proficiency testing described in ISO/IEC Guide 43. To bring the document into harmony with ISO/IEC 17043:2010, the second edition of ISO 13528 was published in September 2015 after a technical revision of the previous version. In short, the new edition focuses more on the statistical design and analysis for proficiency testing schemes [2]. New sections are also added to cover procedures for qualitative proficiency testing schemes and various robust statistical methods.

Participating in proficiency testing schemes has become an essential activity of testing laboratories, and it is also a mandatory requirement for seeking laboratory accreditation according to ISO/IEC 17025. Hence, ISO 13528 is an important reference document for proficiency testing providers and testing laboratories alike, especially in the part related to performance evaluation. For instance, from the perspective of the testing laboratories, the prime concern would be whether the outcome of the performance evaluation through the proficiency testing scheme was sound and valid. On the other hand, the proficiency testing providers would have a concern if the statistical methods they adopted fitted for the intended purposes and fulfilled the requirements of ISO/IEC 17043:2010 [1].

This paper attempted to review and discuss the major changes introduced in the new edition of ISO 13528, particularly in the part on statistical design with respect to the objective of proficiency testing scheme and the considerations on assigned value determination and performance scoring. Moreover, procedures provided in the new edition of ISO 13528 for qualitative proficiency testing schemes and various robust statistical methods would be briefly discussed.

✉ Siu Kay Wong
skwong@govtlab.gov.hk

¹ Government Laboratory, Homantin Government Offices,
Homantin, 88 Chung Hau Street, Kowloon,
Hong Kong, SAR, China

Overview

In the first edition of ISO 13528, a flowchart was given showing the activities requiring the use of statistical methods when operating a proficiency testing scheme, which was followed by provision of various methods for the determination of assigned values and deviation for proficiency assessment, and the calculation of performance statistics. In brief, the design simply depended on whether the assigned values and deviation for proficiency assessment were determined before the proficiency test or not. Different from the previous one, the new edition focuses more on statistical design of proficiency testing schemes with respect to the objectives of the schemes as per the requirements of ISO/IEC 17043:2010. In this connection, new sections are added in the new edition covering the general requirements for statistical methods, basic model and general approaches for performance evaluation. There are also additional procedures for qualitative proficiency testing schemes and various robust statistical methods.

Statistical design

According to ISO/IEC 17043:2010, the proficiency testing provider shall plan the proficiency testing scheme according to the intended objective and purpose. Basically, the objective of the proficiency testing scheme would then determine the details of the statistical design, including methods for identifying the assigned values and performance evaluation. However, in setting the statistical design, there should be considerations on the basic model of measurement and statistical assumptions involved. For ease of illustration, a chart showing the flow is given in Fig. 1.

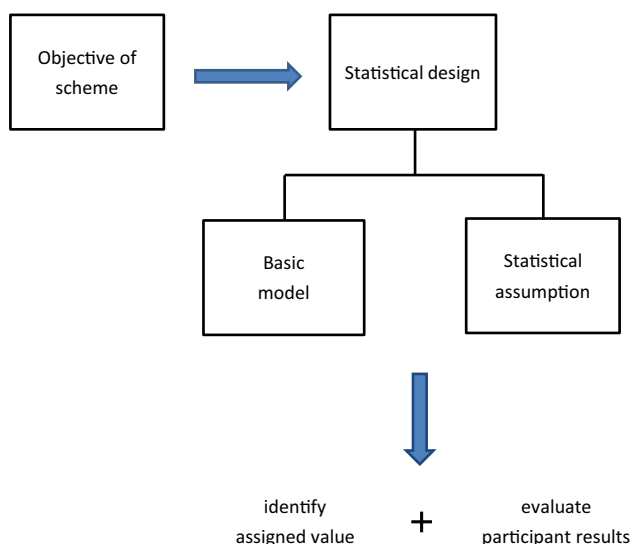


Fig. 1 Flow of statistical design for proficiency testing schemes

Objective of scheme

The first edition of ISO 13528 did not mention much about the objective of proficiency testing scheme and its relationship with statistical design. This is because the aim of the previous edition was to provide estimates of laboratory bias of participants through the use of data from the proficiency tests. Among others, consensus values from participant results were commonly taken as the assigned values as this was the simplest approach though there were still concerns and debates on the appropriateness of this approach under some specific conditions [3, 4]. However, the concept in the new edition of ISO 13528 is to evaluate the fitness of the participant's result which was obtained in the same way as routine laboratory results. Hence, the new edition highlights that proficiency testing schemes might have different objectives and this would lead to different statistical design. Table 1 summarizes the common objectives quoted in the new edition and respective design considerations.

From the perspective of the testing or calibration laboratories, this would allow them to choose schemes according to their needs. For instance, testing laboratories may wish to have their performance be evaluated against those of their peers or a specific group, for example, laboratories accredited for the same test. To this end, they might prefer proficiency testing schemes with objective simply to compare individual participant results with the combined results from participants in the same round. Obviously, for this type of proficiency testing schemes, the assigned value should be determined from the consensus of a specific group or all of the participants in the same round. In this case, there may be no need to have debate on whether it is appropriate using the consensus value as the assigned value for the proficiency testing scheme concerned. Of course, proficiency testing providers have to use appropriate statistical methods to determine the assigned values in a bid to ensure a reliable performance evaluation afterward. In this connection, the new edition of ISO 13528 provides additional information and procedures on various robust statistical methods.

Besides, the objective of a proficiency testing scheme could aim at comparing participant results against reference values which are determined independently, such as by formulation of proficiency test items, appropriate certified reference material, a single laboratory obtained using a reference method in a manner that the value is metrologically traceable to the certified value of an appropriate certified reference material, or an interlaboratory comparison study with expert laboratories as described in ISO Guide 35 [5] for use of interlaboratory comparisons to characterize a certified reference material. These means

Table 1 Common objectives of proficiency testing schemes

Case	Objective of proficiency testing scheme	Design consideration required
1	Compare a participant's result against a predetermined reference value and within limits that are specified before the round begins	Methods for obtaining an externally defined reference value, setting limits and scoring
2	Compare participant results with combined results from a group in the same round and limits that are specified before the round begins	Calculation of an assigned value as well as methods for setting limits and scoring
3	Compare participant results with combined results from a group in the same round and limits determined by the variability of participant results	Calculation of an assigned value and an appropriate measure of dispersion as well as the method of scoring
4	Compare participant results with the assigned value using the participant's own measurement uncertainty	How the assigned value and its uncertainty are to be obtained and how participant measurement uncertainties are to be used in scoring
5	Compare the performance of different measurement methods	Consider the relevant summary statistics and procedures to calculate them

would ensure that the reference value obtained could provide an accurate and reliable estimate of the true value of the measurand for the purpose of performance evaluation. Of course, more resources and efforts are expected for this type of proficiency testing schemes.

Though it is not quite common, the objective of a proficiency testing scheme could be to compare the performance of different measurement methods especially for situations where there would be significant variation among participant results obtained by different measurement methods.

Model and assumption

In the discussion on statistical design, the new edition supplements with general considerations on model and statistical assumptions which form the basis of performance evaluation. For instance, the new edition quotes that, for quantitative results in schemes where a single result is reported for a given test item, the basic model is given in Eq. (1).

$$x_i = \mu + \varepsilon_i \quad (1)$$

with x_i = proficiency test result from participant i ; μ = true value for the measurand; ε_i = measurement error for participant i , distributed according to a relevant model.

The basic model implies that the true value for the measurand could be estimated from the competent participant results. Ideally, if the measurement errors averaged out among the participants in the same round, i.e., $\varepsilon_i \sim N(0, \sigma^2)$, the mean of the participant results would be a close estimate of the true value of the measurand. When the population of x_i was “contaminated” with outliers or erroneous results, robust statistical methods may have to be used to obtain a more reliable consensus value for the

subsequent performance evaluation. As the new edition puts it, for most common analysis techniques, the set of results from competent participants would be approximately normally distributed, or at least unimodal and reasonably symmetric. Of course, it is not uncommon to have distribution of results from competent participants mixed with “contaminated” results from incompetent participants or participants who did not understand the instructions. Considering that the performance evaluation generally relies on the assumption of normality for the distribution of participant results, the new edition hence specifies the need to verify, at least visually, the distribution of participant results against the statistical assumption. This could be done by examining the histogram or kernel plot of the participant results as suggested in the new edition.

However, as the assumption on distribution of participant results would ultimately affect the performance evaluation for the participants, the new edition remarks that the proficiency testing provider has to state the reasons for any statistical assumptions and demonstrate that the assumptions are reasonable. In fact, this is a requirement stipulated in ISO/IEC 17043:2010.

Performance evaluation

According to the new edition of ISO 13528, there are three general approaches for evaluating performance in a proficiency testing scheme, i.e., by comparison with externally derived criteria, other participants' results or claimed measurement uncertainty. Expectedly, the approach adopted should be in agreement with the objective of the proficiency testing scheme. For instance, if the performance is to be evaluated by comparison with other participants, the robust mean of participant results could be

used as the assigned value and the deviation for proficiency assessment may be a predefined allowance for measurement error or a robust standard deviation of participant results.

Regarding scoring methods, z -scores were commonly used where the evaluation did not involve the measurement uncertainties of the participant results. However, if the objective of the scheme was to compare participant results with the assigned value using the participant's own measurement uncertainty, zeta scores (ζ) or E_n scores should be used instead. However, the new edition remarks that a ζ score can be interpreted only as a test of whether the participant's measurement uncertainty is consistent with particular observed deviation and cannot be interpreted as an indication of the fitness for purpose of a particular participant's result. Rather, the fitness for purpose should be judged separately by the participant or by respective accrediting body though comparing the deviation with the target uncertainty as agreed between the testing laboratory and its client.

Reporting of uncertainty data

With a view to improving participants' understanding of measurement uncertainty and its evaluation, the new edition encourages proficiency testing providers to ask participants to report the uncertainty of results in proficiency testing even though uncertainty data are not going to be used in scoring. Among others, this practice should at least be able to achieve the following purposes:

- (a) accreditation bodies can assure that participants are reporting uncertainties that are consistent with their scope of accreditation;
- (b) participants can review their reported uncertainty along with those of other participants, to assess consistency (or not) and therefore gain an opportunity to identify whether the uncertainty is not counting all relevant components, or is over-counting some components; and
- (c) proficiency testing can be used to confirm claims of uncertainty, and this is the easiest when the uncertainty is reported with the result.

If uncertainties are obtained from participants, screening limits on uncertainty could also be derived for participants to identify aberrant uncertainties, for instance, taking uncertainty of the assigned value and 1.5 times the time robust standard deviation of participant results as the lower and upper limits, respectively, for screening reported participants' uncertainties. Certainly, the discussion and suggested procedures on this topic

would help further enhance the functions of proficiency testing schemes.

Qualitative proficiency testing schemes

Apart from those for quantitative analysis, proficiency testing providers and testing laboratories alike would also be interested to have guidance on proficiency testing schemes for qualitative analysis as well, which was not available in the first edition of ISO 13528. Considering that a large amount of proficiency testing occurs for properties that are measured or identified on qualitative scales, the new edition includes guidance on the design and analysis of various types of qualitative testing schemes; for example, schemes require reporting on a nominal scale where the property value has no magnitude.

Similar to that for quantitative analysis, the guidance on qualitative proficiency testing schemes touches on three basic stages, viz. design, value assignment and performance evaluation. First of all, the scheme design would depend on the type of data collected from participants for performance evaluation. For instance, for proficiency testing schemes that report simple, single-valued ordinal results, the proficiency testing provider should consider providing two or more test items per round or requesting the results of a number of replicated observations on each proficiency test item with the number of replicates specified in advance. This would help provide additional information on the nature of errors and also allow more sophisticated scoring of proficiency testing performance.

Depending on the nature of the proficiency test items, the assigned values for qualitative proficiency testing schemes may be assigned by expert judgement; by use of reference materials as proficiency test items; from knowledge of the origin or preparation of the proficiency test item; or using the mode or median of participant results (for ordinal values only).

For proficiency testing schemes in which expert opinion is essential either for value assignment or for assessment of participant performance, the new edition recommends assembling a panel of appropriately qualified experts, which allows discussion and debate required to achieve consensus on appropriate assignment. Also, the proficiency testing provider may choose to circulate the proficiency test items "blind" to different members for the expert panel to assure consistency of diagnosis, or carry out periodic exercise to evaluate the agreement among the panel. Any significant disagreement among the panel should be recorded in the report for the round. If the panel cannot reach a consensus for a particular proficiency test item, the proficiency testing provider may consider an alternative method of value assignment. If that is not appropriate, the

proficiency test item should not be used for performance assessment of participants.

For performance evaluation, the new edition expects that one or more individual experts will review each participant report for each proficiency test item and allocate a performance mark or score. For instance, results that exactly match the assigned value are marked as acceptable and given a corresponding score while results that do not exactly match the assigned value are given a score that depends on the nature of the mismatch. Where multiple replicates are reported for each proficiency test item or where multiple proficiency test items are provided to each participant, the proficiency testing provider may calculate and use combined performance scores or score summaries in performance assessment. Combined performance scores or score summaries may be calculated as, for example, simple sum of performance scores across all proficiency test items.

For further information, the new edition of ISO 13528 includes discussion on some typical examples like proficiency testing in forensic comparisons. Also, an example for the analysis of ordinal data for a proficiency testing scheme is provided for reference purposes.

Statistical data analysis

Proficiency testing schemes inevitably involve statistical data analysis, for example, to determine the assigned values and standard deviation for proficiency assessment from the participants results. However, as most proficiency testing data sets include a proportion of results that are unexpectedly distant from the majority, the capability of the adopted statistical technique regarding robustness to contaminated populations would be critical to achieving a reliable performance evaluation for participants in a proficiency testing scheme. In view of this, the first edition of ISO 13528 included statistical procedures, i.e., Algorithm A, for the determination of robust values of the average and standard deviation of the data. However, there are situations, for example, where the proficiency testing scheme has few participants or has a large portion of results that can be discrepant and the Algorithm A may not be an efficient approach. To address this issue, the new edition of ISO 13528 discusses and provides procedures for various robust statistical methods with different capabilities regarding robustness to outliers and simplicity of application. In particular, efficiency, breakdown point and sensitivity to minor modes of various robust statistical methods were discussed and compared. The robust statistical methods presented in the new edition, in order of simplicity, are sample mean/standard deviation, median/scaled median absolute deviation (MAD), median/

normalized interquartile range (nIQR), Algorithm A and Q_n/Q methods. As noted, the order of simplicity is approximately inversely related to efficiency because the more complex methods tend to have been developed in order to improve efficiency. For instance, though the procedures involved are quite complicated, the Q_n/Q methods display a very good resistance to minor modes and are particularly useful for situations where over 20 % of results can be discrepant.

Of course, though the choice of statistical methods is the responsibility of the proficiency testing provider, it would be beneficial if the participating laboratories would have an understanding about the procedures of the adopted statistical method and the reasons behind why it was adopted for that particular proficiency testing scheme. Indeed, only with such knowledge and considerations could it be judged whether the performance evaluation would be reliable.

Other additional information

Apart from the topics discussed above, the new edition of ISO 13528 also includes discussion on handling of censored results and outlier techniques for individual participant results. For testing of homogeneity of proficiency test items, the new edition suggests expanding the assessment criterion to allow for the actual sampling error and repeatability in the homogeneity check. Also, there are actions proposed for the proficiency testing provider consideration if the criteria for sufficient homogeneity are not met. Regarding procedures for checking stability, the new edition of ISO 13528 provides more detailed considerations with reference to ISO Guide 35. In particular, there are discussions on effects of transportation on proficiency testing items and options if the assessment criteria are not met.

Conclusion

As noticed, the revision of ISO 13528 introduced essential amendments which bring the document into harmony with ISO/IEC 17043:2010, which include a copious amount of new information and update to the first edition. Among others, the new edition stresses on the need of setting the statistical design with respect to the objective of the proficiency testing scheme. This is important as the determination of the assigned value and performance scoring should depend on the objective of the scheme with an appropriate statistical design. Moreover, the new edition includes procedures for qualitative proficiency testing schemes and supplements with discussion and procedures for various robust statistical methods. All in all, the new edition of ISO 13528 offers the necessary guidance and information not

only for the proficiency testing providers to follow but also for the participating laboratories to have a better understanding about the operation of a proficiency testing scheme, particularly on the part of performance evaluation.

References

1. ISO/IEC 17043 (2010) Conformity assessments—general requirements for proficiency testing. International Organization for Standardization (ISO), Geneva
2. ISO 13528 (2015) Statistical methods for use in proficiency testing by interlaboratory comparisons. International Organization for Standardization (ISO), Geneva
3. Kuselman I, Fajgelj A (2010) IUPAC/CITAC guide: selection and use of proficiency testing schemes for a limited number of participants—chemical analytical laboratories. *Pure Appl Chem* 82:1099–1135
4. Emons H (2013) Exploring interlaboratory comparisons. *Accred. Qual Assur* 18:267–268
5. ISO Guide 35 (2006) Reference materials—general and statistical principles for certification. International Organization for Standardization (ISO), Geneva