TECHNICAL SPECIFICATION

ISO/TS 16489

First edition
2006-05-15

Water quality — Guidance for establishing the equivalency of results

Qualité de l'eau — Lignes directrices pour la création de l'équivalence des résultats

Reference number
ISO/TS 16489:2006(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In other circumstances, particularly when there is an urgent market requirement for such documents, a technical committee may decide to publish other types of normative document:

— an ISO Publicly Available Specification (ISO/PAS) represents an agreement between technical experts in an ISO working group and is accepted for publication if it is approved by more than 50 % of the members of the parent committee casting a vote;

— an ISO Technical Specification (ISO/TS) represents an agreement between the members of a technical committee and is accepted for publication if it is approved by 2/3 of the members of the committee casting a vote.

An ISO/PAS or ISO/TS is reviewed after three years in order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn. If the ISO/PAS or ISO/TS is confirmed, it is reviewed again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TS 16489 was prepared by Technical Committee ISO/TC 147, *Water quality*, Subcommittee SC 2, *Physical, chemical and biochemical methods*.

Introduction

The methods referred to in this Technical Specification can comprise a standard or reference method, the results of which are to be compared with results generated by an alternative, perhaps more simple, method. Alternatively, a comparison of results produced by an old established method and those produced by a new more modern technique can be undertaken. The methods can be laboratory based or undertaken “on-site” where the samples are taken.

No indication is given to confirm whether either one of the two methods, in terms of bias, is better or worse than the other method, only that the results produced by both methods are considered equivalent or not, in terms of the calculated means, standard deviations and variances. The procedures described are not to be used for, and do not apply to, situations to establish whether two methods can be shown to be equivalent. The procedures apply only to demonstrating equivalency of results.

Since standard deviations and means can vary with concentrations, especially where concentrations vary over several orders of magnitude, the procedures described in Clauses 7 to 9 are only applicable to samples containing a single level of concentration. It would be necessary to repeat the procedures for each concentration level if different concentration levels are encountered, and it is shown that standard deviations and means vary over these concentration levels. It might be that the demonstration of equivalence can only be achieved over relatively small concentration ranges. For multiple concentration levels, the procedures described in Clause 10 might be applicable. In addition, the laboratory will need to show that both methods are suitable and appropriate for the sample matrix and the parameter under investigation, including the level of concentration of the parameter. Also, the experimental data obtained in the comparison of results should reflect the specific application for which equivalence is questioned, as different matrices can lead to different results with the two methods.

Throughout this Technical Specification, it is assumed that results are obtained essentially under repeatability conditions, but it is recognized that this will not always be so. Hence, where appropriate, identical samples are analysed by the same analyst using the same reagents and equipment in a relatively short period of time. Furthermore, a level of confidence of 95 % is assumed. The statistical tests described in this Technical Specification assume that the data to be compared are independent and normally distributed in a Gaussian manner. If they are not, the data might not be suitable for the statistical treatments described and additional data might need to be collected.

The power of the statistical test is greatly enhanced when sufficient data are available for comparisons; i.e. when the numbers of degrees of freedom are available to enable a meaningful interpretation to be made. However, it is recognized that a statistically significant difference might not necessarily infer an important or meaningful difference, and a personal judgement should be made on whether a statistically significant difference is important or meaningful and relevant. Alternatively, a statistical test might not be sufficiently powerful to be able to detect a difference that from a practical point of view could be regarded as important or meaningful.

To aid the analyst, advice is provided as to which clause (and corresponding annex) is applicable to the circumstances surrounding the data that have been generated. It is recognized that when results are compared they can have been generated under a variety of different conditions.

Water quality — Guidance for establishing the equivalency of results

1 Scope

This Technical Specification describes statistical procedures to test the equivalency of results obtained by two different analytical methods used in the analysis of waters. This Technical Specification is not applicable for establishing whether two methods can be shown to be equivalent. The procedures given in this Technical Specification are only applicable to demonstrating the equivalency of results.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 5725-2, *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*

NOTE A practical guidance document to assist in the use of ISO 5725-2 has been published: see ISO/TR 22971.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply:

3.1
precision
closeness of agreement between independent test results obtained under repeatability conditions

NOTE 1 Precision depends only on the distribution of random errors and does not relate to the true, specified or accepted value.

NOTE 2 Measurement of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results. Less precision is reflected by a larger standard deviation.

NOTE 3 “Independent test results” means results obtained in a manner not influenced by any previous result on the same sample. Quantitative measurements of precision depend critically on stipulated conditions.

3.2
repeatability conditions
conditions where independent test results are obtained with the same method on identical test samples in the same laboratory, by the same operator, using the same reagents and equipment within short intervals of time

3.3
analytical method
unambiguously written procedure describing all details required to carry out the analysis of the determinand or parameter, namely: scope and field of application, principle and/or reactions, definitions, reagents, apparatus, analytical procedures, calculations and presentation of results, performance data and test report

4 Overview of the different approaches

Where a sample is analysed in replicate using two methods, then the procedures described in Clause 7 and Annex B may be used. The results should, ideally, be generated by a single analyst, however, it is recognized that different analysts can be involved.

The procedures described in Clause 8 and Annex C might be applicable where, over a period of time, samples are analysed by different analysts using a particular method and these results are compared with results generated using an alternative method that is carried out by one or more analysts. In this case, however, the assumption of repeatability will not be applicable.

Where different analysts are involved in the generation of data, the procedures described in Clause 9 and Annex D may be used. In these cases, the assumption of repeatability will not be applicable. Where identical samples are analysed by one or more analysts using two different methods, the procedures described in Clause 10 and Annex E might be more appropriate. This might be applicable where the same or different concentration levels are indicated.

5 Amount of data

The approach described in this Technical Specification demonstrates the importance that the power of the significance tests lies in the amount of data available as well as the quality (spread) of the data. Throughout this Techncial Specification, it is assumed that the level of confidence is established at 95 %. This might represent a degree of acceptability that is insufficient for certain purposes. This would mean that individual circumstances would merit individual consideration as to whether this Technical Specification, in terms of the confidence level used, should be applied. Confidence levels of 99 % or higher might be, in certain circumstances, more appropriate. In addition, where a statistically significant difference has been suggested by a statistical analysis of the data, there is always a need to question whether this difference is important or relevant, in terms of its suitability and fitness for purpose, and not in terms of its statistical meaning or understanding. This judgement should be based on whether the analytical results are fit for their intended purpose.

For example, with large amounts of data, it is possible to conclude that there is a statistically significant difference between 50,1 and 50,2. Whether this difference is important or meaningful is another matter when deciding on the suitability of the method.

Before any statistical treatment is undertaken, it is always useful to plot a graph of the data. This will provide a visual display of the results, an inspection of which should reveal the amount and quality of data available for comparison. In this way, the number of results and the spread (or range) of the data is easily observed. Figures F.1 to F.6 (Annex F) show illustrative examples of the type of plots that can be produced and the interpretations that can be concluded. Figures F.1 to F.3 show the arithmetic means of the results from a series of determinations undertaken in comparative exercises of two methods and the associated interpretations. Figures F.4 to F.6 show the spread or range of results from a series of determinations and possible interpretations.

From the data, the arithmetic mean (average) $\bar{x}$ of a number, n, of determinations or measurements, x_i, and the standard deviation, s, of numerous repeated determinations obtained under repeatability conditions, are calculated from Equations (1) and (2):

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} \tag{1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{i=n} x_i^2 - \frac{\left(\sum_{i=1}^{i=n} x_i\right)^2}{n}}{n-1}} \tag{2}$$

The square of the standard deviation is known as the variance, namely, s^2.

6 Data comparisons

When the results from two methods are compared, different situations will arise depending upon the circumstances surrounding the manner in which the results are determined. Hence, the comparison will differ for different situations. By way of example, Clauses 7 to 10 describe the different approaches that can be encountered when sets of data are to be compared. In addition, since the comparisons undertaken in this Technical Specification are used to establish whether a difference between sets of data exists, rather than to determine whether one set of data is superior to another, then a two-sided test is carried out, rather than a one-sided test.

Data comparisons can be further complicated by the inclusion of outlier tests to establish whether sets of data contain values that are considered significantly different from the rest of the data. A number of different outlier tests are available and some of these are described in more detail in ISO 5725-2. Other outlier tests may also be used, for example see Annex E. Further consideration of, and the need for, outlier tests are not considered in this Technical Specification but will need to be taken into consideration.

The example comparisons and information contained in Figures F.1 to F.6 and Annexes B to E are for illustrative purposes only. Suitable computer software might be available to facilitate the numerical calculations. In addition, the examples shown are based on limited data to highlight the manner in which the calculations were carried out. They are not presented as actual data comparisons. In reality, many more results would be required before calculations of this type are undertaken. Schematic diagrams outlining the procedures that can be undertaken are shown in Figures G.1 and G.2 in Annex G.

Samples for analysis should be taken using procedures given in relevant International Standards appropriate to the parameter being analysed.

7 Comparison of arithmetic means of two independently obtained sets of data

Under repeatability conditions, analyse a sample in replicate using the two methods. The number of replicate determinations or measurements carried out with each method can be different, but for both methods should be sufficient to provide confidence in the statistical treatment that follows. This may involve 6 to 10 or more repeat determinations. For example, for the analytical method, method i, the following determinations can be obtained, namely x_1, x_2, x_3, $x_4 \ldots x_{n-1}$ and x_n. For the alternative analytical method, method j, the following determinations can be obtained, namely y_1, y_2, $y_3 \ldots y_{m-1}$ and y_m. From these values the corresponding means, standard deviations and variances are calculated, $\bar{x}$, $\bar{y}$, s_i, s_j, s_i^2 and s_j^2 respectively.

To ascertain whether the precision or spread of data (in terms of the variances s_i^2 and s_j^2) obtained from the two methods differ statistically, a statistical F-test should be carried out. This statistical test will show whether there is a statistically significant difference between the two variances. The F-value calculated (F_{calc}) should then be compared with the tabulated or theoretical F-value (F_{tab}) obtained for the corresponding amount of data; i.e. number of degrees of freedom, at the stated level of confidence required, in this case 95 % (see Table A.1). If F_{tab} is less than F_{calc}, then it can be concluded that there is a statistically significant difference between the two variances; i.e. s_i^2 and s_j^2 are not the same and, hence, cannot be regarded as being equivalent.

Under these circumstances, the variances should not be combined to form a single variance value. The method exhibiting the smaller variance is the more precise of the two methods.

If F_{tab} is greater than F_{calc}, then it can be concluded that there is no statistically significant difference between the two variances; i.e. s_i^2 and s_j^2 can be regarded as being similar and, hence, can be regarded as being equivalent. Under these circumstances, the precision of the results generated by both methods can be regarded as being equivalent.

F_{calc} should be calculated as follows:

$$F_{calc} = \frac{s_i^2}{s_j^2} \quad \text{or} \quad F_{calc} = \frac{s_j^2}{s_i^2} \tag{3}$$

The equation is always arranged so that a value greater than 1 is obtained.

If no statistically significant difference is indicated for the variances, i.e. if F_{tab} is greater than F_{calc}, then the spread of results from both methods can be regarded as being similar. In such a case, the results from both methods can be combined to produce a pooled or combined standard deviation, s_c, according to Equation (4):

$$s_c = \sqrt{\frac{s_i^2(n-1)+s_j^2(m-1)}{n+m-2}} \tag{4}$$

To ascertain if the arithmetic means, $\bar{x}$, $\bar{y}$, obtained for both methods differ statistically, a t-test should be carried out. This test will show whether there is a statistically significant difference between the two means. The t-value calculated (t_{calc}) should then be compared with the tabulated or theoretical t-value (t_{tab}) obtained for the corresponding amount of data; i.e. number of degrees of freedom, at the stated level of confidence required, in this case 95 % (see Table A.2). If t_{tab} is less than t_{calc}, then it can be concluded that there is a statistically significant difference between the two arithmetic means; i.e. $\bar{x}$ and $\bar{y}$ are not the same, and hence cannot be regarded as being equivalent.

If t_{tab} is greater than t_{calc}, then it can be concluded that there is no statistically significant difference between the two means; i.e. $\bar{x}$ and $\bar{y}$ can be regarded as being similar and, hence, can be regarded as being equivalent. Under these circumstances, the bias of the results generated by both methods can be regarded as being equivalent.

t_{calc} should be calculated as follows:

$$t_{calc} = \frac{\left|(\bar{x}-\bar{y})\right|}{s_c\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} \tag{5}$$

Using these tests, it can be concluded that the precision and bias of the results generated for both methods might or might not be similar. Only if the precision (in terms of s_i^2 and s_j^2) and bias (in terms of $\bar{x}$ and $\bar{y}$) of both sets of results show no statistically significant difference can the results be considered equivalent.

An example of this approach is shown in Annex B.

The use of these statistical tests can also indicate whether the method performance capabilities change significantly over periods of time from those originally established. In these instances, it might be that analytical quality control data can be used and compared over the two time periods rather than considering the data being generated by two different methods.

8 Comparison of population and sample arithmetic means

Over a long period of time, a method might be used by different analysts which provides sufficient information to be established, for example on the overall arithmetic mean, μ, of quality control samples. If a different method is then used by a number of analysts and information gathered on its performance, for a (small) number of determinations, n, the arithmetic mean, $\bar{x}$, and standard deviation, s, can be calculated from results obtained using the new method.

To ascertain whether the results from the new method differ statistically from the results obtained by the old method, a t-test should be carried out. This test will show whether there is a statistically significant difference between the two means, μ and $\bar{x}$. The t-value calculated (t_{calc}) should then be compared with the tabulated or theoretical t-value (t_{tab}) obtained for the corresponding amount of data; i.e. number of degrees of freedom, at the stated level of confidence required (see Table A.2). If t_{tab} is less than t_{calc}, then it can be concluded that there is a statistically significant difference between the two arithmetic means; i.e. μ and $\bar{x}$ are not the same, and hence, cannot be regarded as being equivalent.

If t_{tab} is greater than t_{calc}, then it can be concluded that there is no statistically significant difference between the two means; i.e. μ and $\bar{x}$ can be regarded as being similar, and hence, can be regarded as being

equivalent. Under these circumstances, the bias of the results generated by both methods can be regarded as being equivalent.

On this occasion, t_{calc} should be calculated as follows:

$$t_{calc} = \frac{\left|(\bar{x} - \mu)\right|}{\frac{s}{\sqrt{n}}} \tag{6}$$

An example of this approach is shown in Annex C.

As well as demonstrating the equivalency of results, this test can also be used to ascertain if a method that is used and exhibits a certain bias is deemed acceptable when compared with a target bias value. For example, a method exhibiting a bias of, say, 10,5 % might or might not be statistically acceptable when compared with a target bias value of, say, 10 %. Hence, the actual method performance can be compared to a required level of performance.

9 Analysis of variance

When a new method is proposed, a number of different analysts might be used to generate results or performance data to demonstrate its capability. Under these circumstances, when repeat determinations are made, there will always be some variability in the results and it can be difficult to ascertain if real differences exist between the different sets of data produced by the different analysts. One way this could be undertaken would be to carry out repeated t-tests, as described in Clause 7. Repeated use of this test to compare all combinations of data sets, however, increases the probability of making erroneous conclusions. An easier way is to carry out an analysis of variance (ANOVA) test. This test will help to ascertain whether there are statistically significant differences between the sets of data generated by the different analysts. In other words, the ANOVA test is used to determine the statistical significance of differences in the arithmetic means of different sets of data. The data should be arranged as indicated in Table 1, and then an F-test should be carried out. The F-value calculated (F_{calc}) should then be compared with the tabulated or theoretical F-value (F_{tab}) obtained for the corresponding amount of data i.e. number of degrees of freedom, at the stated level of confidence required, in this case 95 %, (see Table A.1).

Table 1 — Statistical significance of differences in the arithmetic means of different sets of data

Replicate determinations	Analysts					
	1	2	3	i	$p-1$	p
1	x_{11}	x_{21}	x_{31}	x_{i1}	$x_{(p-1)1}$	x_{p1}
2	x_{12}	x_{22}	x_{32}	x_{i2}	$x_{(p-1)2}$	x_{p2}
k	x_{1k}	x_{2k}	x_{3k}	x_{ik}	$x_{(p-1)k}$	x_{pk}
$n-1$	$x_{1(n-1)}$	$x_{2(n-1)}$	$x_{3(n-1)}$	$x_{i(n-1)}$	$x_{(p-1)(n-1)}$	$x_{p(n-1)}$
n	x_{1n}	x_{2n}	x_{3n}	x_{in}	$x_{(p-1)n}$	x_{pn}

Equations (7) to (9) should then be used to calculate the test statistics.

$$A = \sum_{i=1}^{i=p} \frac{\left(\sum_{k=1}^{k=n} x_{ik}\right)^2}{n} \tag{7}$$

$$B = \sum_{i=1}^{i=p} \sum_{k=1}^{k=n} x_{ik}^{\,2} \tag{8}$$

$$C = \frac{\left(\sum_{i=1}^{i=p} \sum_{k=1}^{k=n} x_{ik} \right)^2}{N} \tag{9}$$

where the total number of replicates $N = np$.

If the number of replicates for each analyst is not the same, then Equation (10) should be used instead of Equation (7) to calculate the test statistic

$$D = \sum_{i=1}^{i=p} \frac{\left(\sum_{k=l}^{k=n_i} x_{ik} \right)^2}{n_i} \tag{10}$$

where n_i is the number of replicates determined by each analyst.

In addition, the value of N should be calculated accordingly.

To ascertain whether the sets of data differ statistically, an ANOVA table of results should be calculated as shown in Table 2:

Table 2 — ANOVA results

Source of variation	Sum of squares, S	Degrees of freedom	Mean square, M	F_{calc}
Between analysts	$S_1 = A - C$	$p - 1$	$M_1 = S_1/(p - 1)$	M_1/M_0
Within analysts	$S_0 = B - A$	$N - p$	$M_0 = S_0/(N - p)$	
Total	$S_1 + S_0 = B - C$	$N - 1$		
NOTE A, B and C are defined in Equations (7) to (9).				

If F_{tab} is less than F_{calc}, then it can be concluded that there is a statistically significant difference between the sets of data. Hence, one or more of the sets of data is not the same as the remaining sets of data, and hence the whole data set cannot be regarded as being equivalent.

If F_{tab} is greater than F_{calc}, it can be concluded that there is no statistically significant difference between all the sets of data that can be regarded as being similar and hence, equivalent. In such a case, all the results can be combined and the overall mean, standard deviation and variance values calculated. The procedure described in Clause 7 can then be used to ascertain whether the results, in terms of means and variances, obtained for this new method are significantly different from the corresponding values obtained using an existing method.

One disadvantage to this technique is that, if a statistically significant difference is indicated, no information is provided as to which set of data might be different from the remaining sets of data. It might, however, be the case that a graphical plot of the results suggests which set of data is different from the remaining sets. Removing this particular set of data and repeating the procedure might provide confirmation.

An example of this approach is shown in Annex D.