

# COMPARISON OF ALTERNATIVE MEASUREMENT METHODS

## 1. Introduction

- 1.1. This document describes the comparison of the accuracy (trueness and precision) of an analytical method with a reference method. It is based on ISO-5725-6:1994(E), section 8 (Comparison of alternative methods). Where possible the ISO text was taken over and terms used in this document are in accordance with ISO definitions. However the ISO text differs on the following main points from the present text:
- 1.2. In the ISO standard the reference method is an international standard method that was studied in an interlaboratory fashion. This means that the precision ( $\sigma^2$ ) is known. Here we consider the situation in which a laboratory has developed a first method (method A) and validated this, and later on wishes to compare a new method (method B) to the older, already validated method. The former will be referred to as the reference method. Only an estimate of the precision ( $s^2$ ) is available.
- 1.3. This document is meant for use within a single organisation, while the ISO-standard concerns interlaboratory experiments. This means that, either one laboratory will carry out the experiments, or else two laboratories of the same organisation, each laboratory being specialized in one of the methods.
- 1.4. As a consequence of point 1.3 here precision is not investigated under reproducibility conditions. Instead time-different intermediate precision conditions have been considered.
- 1.5. The ISO standard is meant to show that the two methods have similar precision and/or trueness. The present proposal is meant to show that the alternative method is at least as good as the standard method. This means that, in some instances where ISO applies two-sided tests, here one-sided tests are used.

- 1.6. Two different approaches are considered. The first one is based ,as in the ISO document, on the minimal number of measurements required to detect a specified bias between both methods and a specified ratio of the precision of both methods with high probability.

Since this might lead to a number of measurements to be performed that the laboratory considers too large, the second approach starts from a user-defined number of measurements. The probability ( $\beta$ ) is then evaluated that an alternative method which is not acceptable, because it is too much biased and/or not precise enough, will be adopted.

This means that in both approaches an acceptable bias and an acceptable ratio of the precision measures of both methods have to be defined.

- 1.7. The evaluation of the bias is also based on interval hypothesis testing [1] in which the probability of accepting a method that is too much biased is controlled. The bias is considered acceptable if the one-sided 95% upper confidence limit around the estimated absolute bias does not exceed the acceptance limit for the bias.

## **2. Purpose of comparing measurement methods**

The comparison of measurement methods will be required if a laboratory wishes to replace a method which is the recommended or official method in a particular field of application by an alternative method. The latter method should be at least as good (in terms of precision and trueness) as the first method.

## **3. Field of application**

The document describes the comparison of the accuracy (trueness and precision) of two methods at a single concentration level. It is useful for comparisons at up to three concentration levels. Due to the problem with multiple comparisons [2] it should not be used if the methods are to be compared at more than three levels.

## **4. Accuracy experiment**

### **4.1. General requirements**

The procedures for both methods shall be documented in sufficient detail so as to avoid misinterpretation by the participating analysts. No modification to the procedure is permitted during the experiment.

## **4.2. Test samples**

The precision of many measurement methods is affected by the matrix of the test sample as well as the level of the characteristic. For these methods, comparison of the precision is best done on identical test samples. Furthermore, comparison of the trueness of the methods can only be made when identical test samples are used. For this reason, communication between the working groups who conduct the accuracy experiments on each method should be achieved by appointment of a common executive officer.

The main requirement for a test sample is that it shall be homogeneous and stable, i.e., each laboratory shall use identical test samples. If within-unit inhomogeneity is suspected, clear instructions on the method of taking test portions shall be included in the document. The use of reference materials (RMs) for some of the test samples has some advantages. The homogeneity of the RM has been assured and the results of the method can be examined for bias relative to the certified value of the RM. The drawback is usually the high cost of the RM. In many cases, this can be overcome by redividing the RM units. For the procedure for using a RM as a test sample, see ISO Guide 33.

## **4.3. Number of test samples**

The number of test samples used varies depending on the range of the characteristic levels of interest and on the dependency of the accuracy on the level. In many cases, the number of test samples is limited by the amount of work involved and the availability of a test sample at the desired level.

## **4.4. Number of measurements.**

### **4.4.1. Determination of the minimal number of measurements required**

In this approach the minimal number of measurements required, to detect a specified bias between two methods and a specified ratio of the precision of both methods with high probability, is determined.

#### **4.4.1.1. General**

The number of days and the number of measurements per day required for both methods depends on:

- a) precisions of the two methods;
- b) detectable ratio,  $\rho$  or  $\phi$ , between the precision measures of the two methods; this is the minimum ratio of precision measures that the experimenter wishes to detect with high probability from the results of experiments using two methods;

the precision may be expressed as the repeatability standard deviation, in which case the ratio is termed  $\rho$ , or as the square root of the between-day mean squares, in which case the ratio is termed  $\phi$ ;

c) detectable difference between the biases of the two methods,  $\lambda$ ; this is the minimum value of the difference between the expected values of the results obtained by the two methods that the experimenter wishes to detect with high probability from the results of an experiment.

It is recommended that a significance level of  $\alpha = 0.05$  is used to compare precision estimates and to evaluate the bias of the alternative method. The risk of failing to detect the chosen minimum ratio of standard deviations, or the minimum difference between the biases, is set at  $\beta = 0.20$ .

With those values of  $\alpha$  and  $\beta$ , the following equation can be used for the detectable difference:

$$\lambda = (t_{\alpha/2} + t_{\beta}) \sqrt{\frac{(p_A - 1)(s_{tA}^2 + s_{rA}^2/n_A) + (p_B - 1)(s_{tB}^2 + s_{rB}^2/n_B)}{p_A + p_B - 2} \left( \frac{1}{p_A} + \frac{1}{p_B} \right)} \quad (1)$$

where the subscripts A and B refer to method A and method B, respectively.

$t_{\alpha/2}$  : two-sided tabulated t-value at significance level  $\alpha$  and degrees of freedom  $v = p_A + p_B - 2$

$t_{\beta}$  : one-sided tabulated t-value at significance level  $\beta$  and degrees of freedom  $v = p_A + p_B - 2$

$s_t^2$  : estimated variance component between days

$s_r^2$  : estimated repeatability variance component

$p$  : number of days

$n$  : number of measurements within one day

In most cases, the precision of method B is unknown. In this case, use the precision of method A as a substitute to give

$$\lambda = (t_{\alpha/2} + t_{\beta}) \sqrt{\frac{(p_A - 1)(s_{tA}^2 + s_{rA}^2/n_A) + (p_B - 1)(s_{tA}^2 + s_{rA}^2/n_B)}{p_A + p_B - 2} \left( \frac{1}{p_A} + \frac{1}{p_B} \right)} \quad (2)$$

The experimenter should try substituting values of  $n_A$ ,  $n_B$ ,  $p_A$  and  $p_B$  (and the corresponding  $t_{\alpha/2}$  and  $t_{\beta}$ ) in equation (1) or (2) until values are found which are

large enough to satisfy the value of  $\lambda$  chosen (i.e. so that  $\lambda$  computed with eq. 2 is smaller than the stated acceptable  $\lambda$ ).

It is strongly recommended to take  $n_A = n_B$  and  $p_A = p_B$ . In this case eq (2) simplifies to

$$\lambda = (t_{\alpha/2} + t_{\beta}) \sqrt{2 \left( s_{tA}^2 + s_{tA}^2 / n_A \right) / p_A} \quad (3)$$

*NOTE : assuming  $s_A$  to be equal to  $s_B$  is of course a strong assumption since even if  $\sigma_A = \sigma_B$  it is improbable for  $s_A$  to be equal to  $s_B$ . Therefore eqs.(2 and 3) are only approximates which could be further simplified by replacing  $(t_{\alpha/2} + t_{\beta})$  by a constant value. Indeed for  $\alpha=0.05$  and  $\beta=0.20$ ,  $(t_{\alpha/2} + t_{\beta})$  varies between 2.802 ( $\nu = \infty$ ) and 3.195 ( $\nu=8$  i.e.  $p_A = p_B = 5$ ) and therefore a constant value equal to 3 could be used throughout. Equation (3) then becomes:*

$$\lambda = 3 \sqrt{2 \left( s_{tA}^2 + s_{tA}^2 / n_A \right) / p_A}$$

The values of the parameters which are needed for an adequate experiment to compare precision estimates should then be considered. Table 1 shows the minimum ratios of standard deviation for given values of  $\alpha$  and  $\beta$  as a function of the degrees of freedom  $\nu_A$  and  $\nu_B$ .

For repeatability standard deviations

$$\nu_A = p_A (n_A - 1) \text{ and } \nu_B = p_B (n_B - 1)$$

For between-day mean square :

$$\nu_A = (p_A - 1) \text{ and } \nu_B = (p_B - 1)$$

If the precision of one of the methods is well established use degrees of freedom equal to 200 from Table 1.

**Table 1**

Values of  $\rho$  ( $v_A, v_B, \alpha, \beta$ ) or  $\phi$  ( $v_A, v_B, \alpha, \beta$ ) for  $\alpha = 0.05$  and  $\beta = 0.20$

$v_B$	$v_A$																				
	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	50	200
3	5.22	4.41	4.00	3.76	3.60	3.48	3.39	3.32	3.27	3.23	3.19	3.16	3.13	3.11	3.09	3.07	3.05	3.04	2.99	2.89	2.81
4	4.76	3.98	3.59	3.35	3.19	3.08	2.99	2.92	2.87	2.83	2.79	2.76	2.74	2.71	2.69	2.68	2.66	2.65	2.60	2.49	2.42
5	4.51	3.74	3.35	3.12	2.96	2.85	2.77	2.70	2.65	2.60	2.57	2.54	2.51	2.49	2.47	2.45	2.44	2.42	2.37	2.27	2.20
6	4.35	3.59	3.21	2.97	2.82	2.70	2.62	2.55	2.50	2.46	2.42	2.39	2.37	2.34	2.32	2.31	2.29	2.28	2.22	2.12	2.05
7	4.24	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.36	2.32	2.29	2.26	2.24	2.22	2.20	2.19	2.17	2.12	2.02	1.94
8	4.15	3.41	3.03	2.79	2.64	2.53	2.44	2.38	2.32	2.28	2.24	2.21	2.19	2.16	2.14	2.13	2.11	2.10	2.04	1.94	1.86
9	4.09	3.35	2.97	2.74	2.58	2.47	2.38	2.32	2.26	2.22	2.18	2.15	2.13	2.10	2.08	2.07	2.05	2.04	1.98	1.88	1.80
10	4.04	3.30	2.92	2.69	2.53	2.42	2.34	2.27	2.22	2.17	2.14	2.11	2.08	2.06	2.04	2.02	2.00	1.99	1.93	1.83	1.75
11	4.00	3.26	2.88	2.65	2.50	2.38	2.30	2.23	2.18	2.14	2.10	2.07	2.04	2.02	2.00	1.98	1.96	1.95	1.89	1.78	1.70
12	3.97	3.23	2.85	2.62	2.47	2.35	2.27	2.20	2.15	2.10	2.07	2.03	2.01	1.98	1.96	1.95	1.93	1.91	1.86	1.75	1.67
13	3.94	3.21	2.83	2.60	2.44	2.33	2.24	2.17	2.12	2.08	2.04	2.01	1.98	1.96	1.94	1.92	1.90	1.89	1.83	1.72	1.64
14	3.92	3.18	2.80	2.57	2.42	2.30	2.22	2.15	2.10	2.05	2.02	1.98	1.96	1.93	1.91	1.89	1.88	1.86	1.81	1.69	1.61
15	3.90	3.17	2.79	2.55	2.40	2.28	2.20	2.13	2.08	2.03	1.99	1.96	1.94	1.91	1.89	1.87	1.86	1.84	1.78	1.67	1.59
16	3.88	3.15	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.98	1.94	1.92	1.89	1.87	1.85	1.84	1.82	1.76	1.65	1.56
17	3.87	3.13	2.75	2.52	2.37	2.25	2.17	2.10	2.04	2.00	1.96	1.93	1.90	1.88	1.86	1.84	1.82	1.80	1.75	1.63	1.55
18	3.86	3.12	2.74	2.51	2.35	2.24	2.15	2.08	2.03	1.98	1.95	1.91	1.89	1.86	1.84	1.82	1.81	1.79	1.73	1.62	1.53
19	3.84	3.11	2.73	2.50	2.34	2.23	2.14	2.07	2.02	1.97	1.93	1.90	1.87	1.85	1.83	1.81	1.79	1.78	1.72	1.60	1.51
20	3.83	3.10	2.72	2.49	2.33	2.22	2.13	2.06	2.01	1.96	1.92	1.89	1.86	1.84	1.82	1.80	1.78	1.76	1.71	1.59	1.50
25	3.79	3.06	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.92	1.88	1.85	1.82	1.79	1.77	1.75	1.73	1.72	1.66	1.54	1.44
50	3.71	2.98	2.60	2.37	2.21	2.09	2.00	1.93	1.88	1.83	1.79	1.76	1.73	1.70	1.68	1.66	1.64	1.62	1.56	1.43	1.32
200	3.65	2.92	2.54	2.31	2.15	2.03	1.94	1.87	1.81	1.76	1.72	1.69	1.66	1.63	1.60	1.58	1.56	1.55	1.48	1.33	1.19

**4.4.1.2. Example:** Determination of iron in iron ores

*NOTE : the example of ISO has been adapted to the situation where the precision is not known but estimated as  $s^2$  and to the application of a one-sided test for the evaluation of the precision.*

**4.4.1.2.1. Background**

Two analytical methods for the determination of the total iron in iron ore are investigated. An estimate of the precision ( $s_{rA}$  and  $s_{tA}$ ) for method A is available. Both methods are presumed to have equal precision. Therefore

$$s_{rA} = s_{rB} = 0.1\% \text{ Fe}$$

$$s_{tA} = s_{tB} = 0.2\% \text{ Fe}$$

**4.4.1.2 2. Requirements**

$$\lambda = 0.4\% \text{ Fe}$$

$$\rho = \phi = 4$$

The minimum number of days required are computed assuming equal number of days and duplicate analyses per day:

$$p_A = p_B \text{ and } n_A = n_B = 2$$

a) For the trueness requirement :

$$0.4 = (t_{\alpha/2} + t_{\beta}) \sqrt{2(0.2^2 + 0.1^2/2)/p_A}$$

With  $p_A = 5$ ,  $(t_{\alpha/2} + t_{\beta}) = 3.195$  and  $\lambda=0.428$ ; with  $p_A = 6$ ,  $(t_{\alpha/2} + t_{\beta}) = 3.107$  and  $\lambda = 0.381$ . Hence  $p_A = p_B = 6$ .

*NOTE : the use of a constant multiplication factor equal to 3 would also yield  $p_A = p_B = 6p$ .*

b) For the precision requirement :

From Table 1 it can be seen that  $\rho=4$  or  $\phi=4$  is reached when  $v_A = v_B = 4$ .

To compare repeatability standard deviations,  
 $v_A = p_A$  and  $v_B = p_B$ , so  $p_A = p_B = 4$ .

To compare between-day mean squares,

$$v_A = p_A - 1 \text{ and } v_B = p_B - 1, \text{ so } p_A = p_B = 5.$$

#### **4.4.1.2.3. Conclusions**

The minimum number of days required (with two measurements per day) is 6. This means that with this sample size ( $n_A = n_B = 2, p_A = p_B = 6$ ), provided that reliable precision estimates were considered

- the probability that it will be decided that there is a bias when in fact there is none is 5% and at the same time the probability that a true bias equal to 0.4% will go undetected is 20% and
- the probability that it will be decided that the precision measures of both methods are different when in fact they are equal is 5% and at the same time the probability that a true ratio between the precision measures of both methods equal to 4 will not be identified as being different is 20%.

#### 4.4.2. User-defined number of measurements

This approach is based on a user-defined number of measurements. This means that the number of days ( $p$ ) and the number of measurements per day ( $n$ ) are defined by the user. It is however strongly recommended to take  $n_A = n_B = 2$ . In the comparison of the results of method A and method B the probability  $\beta$  that an alternative method which is not acceptable, because it is too much biased and/or not precise enough, will be adopted is then evaluated. This probability depends on:

- a) precisions of the two methods;
- b) detectable ratio,  $\rho$  or  $\phi$ , between the precision measures of the two methods; this is the minimum ratio of precision measures that the experimenter wishes to detect with high probability from the results of experiments using two methods; the precision may be expressed as the repeatability standard deviation, in which case the ratio is termed  $\rho$ , or as the square root of the between-day mean squares, in which case the ratio is termed  $\phi$ ;
- c) detectable difference between the biases of the two methods,  $\lambda$ ; this is the minimum value of the difference between the expected values of the results obtained by the two methods that the experimenter wishes to detect with high probability from the results of an experiment;
- d) the significance level  $\alpha$  and the number of measurements.

#### 4.5. Test sample distribution

The executive officer of the intralaboratory test programme shall take the final responsibility for obtaining, preparing and distributing the test samples. Precautions shall be taken to ensure that the samples are received by the participating analysts in good condition and are clearly identified. The participating analysts shall be instructed to analyse the samples on the same basis, for example, on dry basis; i.e. the sample is to be dried at 105°C for  $x$  h before weighing.

#### 4.6. Participating analyst

The laboratory shall assign a staff member to be responsible for organizing the execution of the instructions of the coordinator. The staff member shall be a qualified analyst. If several analysts could use the method, unusually skilled staff (such as the "best" operator) should be avoided in order to prevent obtaining an unrealistically low estimate of the standard deviation of the method. The assigned staff member shall perform the required number of

measurements under repeatability and time-different conditions. The staff member is responsible for reporting the test results to the coordinator within the time specified.

It is the responsibility of this staff member to scrutinize the test results for physical aberrants. These are test results that due to explainable physical causes do not belong to the same distribution as the other test results.

#### 4.7. Tabulation of the results and notation used

With 2 measurements per day ( $n=2$ ), as recommended, the test results for each method can be summarized as in Table 2 where:

- $p$  is the number of days
- $y_{i1}, y_{i2}$  are the two test results obtained on day  $i$  ( $i = 1, \dots, p$ )
- $\bar{y}_i$  is the mean of the test results obtained on day  $i$   
 $= (y_{i1} + y_{i2}) / 2$
- $\bar{\bar{y}}$  is the grand mean  
 $= \frac{1}{p} \sum_{i=1}^p \bar{y}_i$

**Table 2**  
Summary of test results (e.g. for Method A)

Day	Test results	Mean
1	y <sub>11</sub>	$\bar{y}_1$
	y <sub>12</sub>	
⋮	⋮	⋮
i	y <sub>i1</sub>	$\bar{y}_i$
	y <sub>i2</sub>	
⋮	⋮	⋮
p	y <sub>p1</sub>	$\bar{y}_p$
	y <sub>p2</sub>	
		$\bar{\bar{y}}$

#### 4.8. Evaluation of test results

The test results shall be evaluated as much as possible using the procedure described in ISO 5725-2 and ISO 5725-3.

This includes among others that outlier tests are applied to the day means.

##### 4.8.1. Outlying day means

###### 4.8.1.1. One outlier

The day means are arranged in ascending order.

The single-Grubbs' test is used to determine whether the largest day mean  $\bar{y}_p$  is an outlier. Therefore the Grubbs' statistic  $G$  is computed:

$$G = (\bar{y}_p - \bar{\bar{y}}) / s$$

where 
$$s = \sqrt{\frac{1}{p-1} \sum_{i=1}^p (\bar{y}_i - \bar{\bar{y}})^2}$$

To determine whether the smallest day mean  $\bar{y}_1$  is an outlier compute Grubbs' statistic  $G$  as follows:

$$G = (\bar{\bar{y}} - \bar{y}_1) / s$$

Critical values for Grubbs' test are given in the Appendix I.

If at the 5% significance level  $G \leq G_{crit}$ , no outlier is detected.

If at the 1% significance level  $G > G_{crit}$ , an outlier has been detected. It is indicated by a double asterisk and is not included in further calculations.

If at the 5% significance level  $G > G_{crit}$  and at the 1% significance level  $G \leq G_{crit}$ , a straggler has been detected. It is indicated by a single asterisk and is included in the further calculations unless the outlying behaviour can be explained.

###### 4.8.1.2. Two outliers

If the single-Grubbs' test does not detect an outlier, the double-Grubbs' test is used to determine whether the two largest day means are outliers. Therefore the Grubbs' statistic  $G$  is computed as follows:

$$G = SS_{p-1,p} / SS_0$$

where

$$SS_0 = \sum_{i=1}^p (\bar{y}_i - \bar{\bar{y}})^2$$
$$SS_{p-1,p} = \sum_{i=1}^{p-2} (\bar{y}_i - \bar{\bar{y}}_{p-1,p})^2$$

and

$$\bar{\bar{y}}_{p-1,p} = \frac{1}{p-2} \sum_{i=1}^{p-2} \bar{y}_i$$

To determine whether the two smallest day means are outliers the Grubbs' statistic  $G$  is computed as follows:

$$G = SS_{1,2} / SS_0$$

where

$$SS_{1,2} = \sum_{i=3}^p (\bar{y}_i - \bar{\bar{y}}_{1,2})^2$$

and

$$\bar{\bar{y}}_{1,2} = \frac{1}{p-2} \sum_{i=3}^p \bar{y}_i$$

Critical values for the double-Grubbs' test are also given in the Appendix I. Notice that here outliers or stragglers are detected if the test statistic  $G$  is *smaller* than the critical value. The outliers found are indicated by double asterisk and are not included in further calculations. The stragglers found are indicated by a single asterisk and are included in the further calculations unless the outlying behaviour can be explained.

#### 4.8.2. Calculation of variances

A summary for the calculation of the variances is given in Table 3.

For each test sample, the following quantities are to be computed:

$s_{rA}$	is an estimate of the repeatability standard deviation for method A
$s_{rB}$	is an estimate of the repeatability standard deviation for method B
$s_{I(T)A}$	is an estimate of the time-different intermediate precision standard deviation for method A ( $s_{I(T)A}^2 = s_{tA}^2 + s_{rA}^2$ )
$s_{I(T)B}$	is an estimate of the time-different intermediate precision standard deviation for method B ( $s_{I(T)B}^2 = s_{tB}^2 + s_{rB}^2$ )

**Table 3**  
Calculation of variances

ANOVA table

Source	Mean squares	Estimate of
Day	$MS_D = \frac{n \sum_{i=1}^p (\bar{y}_i - \bar{\bar{y}})^2}{(p-1)}$	$\sigma_r^2 + n\sigma_t^2$
Residual	$MS_E = \frac{\sum_{i=1}^p \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}{p(n-1)}$	$\sigma_r^2$

**Calculation of variances**

- The repeatability variance

$$s_r^2 = MS_E \qquad df = p(n-1)$$

- Variance component between days (between-day variance)

$$s_t^2 = \frac{MS_D - MS_E}{n} \qquad \text{if } s_t^2 < 0 \text{ set } s_t^2 = 0$$

- Time-different intermediate precision (variance)

$$s_{I(T)}^2 = s_r^2 + s_t^2 = \frac{MS_D + (n-1)MS_E}{n}$$

- Variance of the means  $\bar{y}_i$

$$s_{\bar{y}}^2 = \frac{\sum_{i=1}^p (\bar{y}_i - \bar{\bar{y}})^2}{p-1} = \frac{MS_D}{n} = s_t^2 + s_r^2/n = s_{I(T)}^2 - (1-1/n)s_r^2 \qquad df = (p-1)$$

#### 4.9. Comparison between results of method A and method B

The results of the test programmes shall be compared for each level. It is possible that method B is more precise and/or biased at lower levels of the characteristic but less precise and/or biased at higher levels of the characteristic values or vice versa.

##### 4.9.1. Graphical presentation

Graphical presentation of the raw data for each level is desirable. Sometimes the difference between the results of the two methods in terms of precision and/or bias is so obvious that further statistical evaluation is unnecessary.

Graphical presentation of the precision and grand means of all levels is also desirable.

##### 4.9.2. Comparison of precision

*NOTE : since we want to evaluate whether the alternative method is as least as good as the reference method the hypotheses to be tested are  $H_0: \sigma_B^2 \leq \sigma_A^2$ ;  $H_1: \sigma_B^2 > \sigma_A^2$ .*

###### 4.9.2.1. Based on the minimum number of measurements required

###### 4.9.2.1.1. Repeatability

$$F_r = \frac{s_{rB}^2}{s_{rA}^2}$$

If

$$F_r \leq F_{\alpha(v_{rB}, v_{rA})} \quad (4)$$

there is no evidence that method B has worse repeatability than method A;

if

$$F_r > F_{\alpha(v_{rB}, v_{rA})}$$

there is evidence that method B has worse repeatability than method A.

$F_{\alpha(v_{rB}, v_{rA})}$  is the value of the F-distribution with  $v_{rB}$  degrees of freedom associated with the numerator and  $v_{rA}$  degrees of freedom associated with the denominator;  $\alpha$  represents the portion of the F-distribution to the right of the given F-value and

$$v_{rA} = p_A(n_A - 1)$$

$$v_{rB} = p_B(n_B - 1)$$

#### 4.9.2.1.2. Time-different intermediate precision

For the comparison of the time-different precision the number of degrees of freedom associated with the precision estimates is needed. Since these estimates are not directly estimated from the data but are calculated as a linear combination of two mean squares,  $MS_D$  and  $MS_E$  (see Table 3), the number of degrees of freedom are determined from the Satterthwaite approximation [3].

However to avoid the complexity in the determination of the degrees of freedom associated with  $s_{I(T)}^2$ , the comparison of the time-different intermediate precision can be performed in an indirect way by comparing the variance of the day means  $s_{y_i}^2$ , provided that the repeatabilities of both methods are equal ( $\sigma_{rA}^2 = \sigma_{rB}^2$ ) and the number of replicates per day for both methods is equal ( $n_A = n_B$ ).

**Check whether**  $\sigma_{rA}^2 = \sigma_{rB}^2$  ( $H_0 : \sigma_{rB}^2 = \sigma_{rA}^2 ; H_1 : \sigma_{rB}^2 \neq \sigma_{rA}^2$ )

$$F = \frac{s_1^2}{s_2^2}$$

with  $s_1^2$  the largest of  $s_{rA}^2$  and  $s_{rB}^2$

If  $F_F \leq F_{\alpha/2}(v_{r1}, v_{r2})$  there is no evidence that both methods have different repeatabilities.

$F_{\alpha/2}(v_{r1}, v_{r2})$  is the value of the F-distribution with  $v_{r1}$  degrees of freedom associated with the numerator and  $v_{r2}$  degrees of freedom associated with the denominator;  $\alpha/2$  represents the portion of the F-distribution to the right of the given F-value and

$$\begin{aligned}
v_{r1} &= v_{rA} = p_A (n_A - 1) \text{ if } s_{rA}^2 > s_{rB}^2 \\
&= v_{rB} = p_B (n_B - 1) \text{ if } s_{rB}^2 > s_{rA}^2 \\
v_{r2} &= v_{rB} = p_B (n_B - 1) \text{ if } s_{rA}^2 > s_{rB}^2 \\
&= v_{rA} = p_A (n_A - 1) \text{ if } s_{rB}^2 > s_{rA}^2
\end{aligned}$$

*NOTE : The results obtained from the comparison of the repeatabilities in Section 4.9.2.1.1 cannot be used here since a one-tailed F-test has been considered there. A non-significant test, which means that the repeatability of method B is acceptable, does not necessarily imply that the repeatabilities of both methods are equal, the repeatability of method B can be better (smaller) than the repeatability of method A.*

**a) Repeatabilities of both methods are equal and  $n_A = n_B$**

If it can be assumed that the repeatabilities of both methods are equal and if  $n_A = n_B$ , the time-different intermediate precisions are compared by calculating  $F_{I(T)}$  as follows:

$$F_{I(T)} = \frac{s_{I(T)B}^2 - (1 - 1/n_B)s_{rB}^2}{s_{I(T)A}^2 - (1 - 1/n_A)s_{rA}^2} = \frac{s_{tB}^2 + s_{rB}^2/n_B}{s_{tA}^2 + s_{rA}^2/n_A} = \frac{s_{\bar{y}B}^2}{s_{\bar{y}A}^2} \quad (5)$$

If

$$F_{I(T)} \leq F_{\alpha(v_{I(T)B}, v_{I(T)A})}$$

there is no evidence that the time-different intermediate precision of method B is worse than that of method A;

if

$$F_{I(T)} > F_{\alpha(v_{I(T)B}, v_{I(T)A})}$$

there is evidence that the time-different precision of method B is worse than that of method A.

$F_{\alpha(v_{I(T)B}, v_{I(T)A})}$  is the value of the F-distribution with  $v_{I(T)B}$  degrees of freedom associated with the numerator and  $v_{I(T)A}$  degrees of freedom associated with the denominator;  $\alpha$  represents the portion of the F-distribution to the right of the given F-value and

$$\begin{aligned} v_{I(T)A} &= p_A - 1 \\ v_{I(T)B} &= p_B - 1 \end{aligned} \tag{6}$$

***b) Repeatabilities of both methods are not equal or  $n_A \neq n_B$***

If it cannot be assumed that the repeatabilities of both methods are equal or if  $n_A \neq n_B$ , the time-different intermediate precisions are compared by calculating  $F_{I(T)}$  as follows:

$$F_{I(T)} = \frac{s_{I(T)B}^2}{s_{I(T)A}^2} \tag{7}$$

We do not have direct estimates of  $s_{I(T)}^2$ . Indeed, the latter, as follows from Table 3, is a compound variance. The number of degrees of freedom associated with  $s_{I(T)}^2$  is obtained from the Satterthwaite approximation:

$$v_{I(T)} = \frac{\left(s_{I(T)}^2\right)^2}{(MS_D/n)^2/(p-1) + ((n-1)MS_E/n)^2/p(n-1)} \tag{8}$$

If

$$F_{I(T)} \leq F_{\alpha(v_{I(T)B}, v_{I(T)A})}$$

there is no evidence that the time-different intermediate precision of method B is worse than that of method A.

If

$$F_{I(T)} > F_{\alpha(v_{I(T)B}, v_{I(T)A})}$$

there is evidence that the time-different intermediate precision of method B is worse than that of method A.

## 4.9.2.2. Based on a user defined number of measurements

### 4.9.2.2.1. Repeatability

The repeatabilities are compared as described in Section 4.9.2.1.1 but additionally the probability  $\beta$  of not detecting a difference which in reality is

equal to  $\rho$  is computed

Therefore if  $F_r \leq F_{\alpha}(v_{rB}, v_{rA})$  (see eq. (4)) calculate:

$$F_{1-\beta}(v_{rB}, v_{rA}) = \frac{F_{\alpha}(v_{rB}, v_{rA})}{\rho^2}$$

$$F_{\beta}(v_{rA}, v_{rB}) = 1/F_{1-\beta}(v_{rB}, v_{rA})$$

and find from an F-table the probability  $\beta$  that  $F \geq F_{\beta}(v_{rA}, v_{rB})$

$F_{\alpha}(v_{rB}, v_{rA})$  is the value of the F-distribution with  $v_{rB}$  degrees of freedom associated with the numerator and  $v_{rA}$  degrees of freedom associated with the denominator;  $\alpha$  represents the portion of the F-distribution to the right of the given F-value and

$$v_{rA} = p_A (n_A - 1)$$

$$v_{rB} = p_B (n_B - 1)$$

$F_{\beta}(v_{rA}, v_{rB})$  is the value of the F-distribution with  $v_{rA}$  degrees of freedom associated with the numerator and  $v_{rB}$  degrees of freedom associated with the denominator;  $\beta$  represents the portion of the F-distribution to the right of the given F-value.

**Example:**  $\rho = 2$                        $p_A = p_B = 7$                        $n_A = n_B = 2$

therefore  $\rho^2 = 4$                        $v_A = v_B = 7$

$$F_{\alpha(7,7)} = 3.77$$

$$F_{1-\beta(7,7)} = 3.77 / 4 = 0.9425$$

$$F_{\beta(7,7)} = 1 / 0.9425 = 1.06$$

$\rightarrow \beta = 47\%$

This means that if in reality  $\sigma_{rB}^2 = 4\sigma_{rA}^2$  the probability that this difference will not be detected is as large as 47%.

#### 4.9.2.2.2. Time-different intermediate precision

The time-different intermediate precisions are compared as described in Section 4.9.2.1.2 but additionally the probability  $\beta$  of not detecting a difference which in reality is equal to  $\phi$  is computed.

##### a) Repeatabilities of both methods are equal and $n_A = n_B$

Proceed as described in Section 4.9.2.1.2a.

If  $F_{I(T)} \leq F_{\alpha(v_{I(T)B}, v_{I(T)A})}$  to evaluate  $\beta$ , calculate:

$$F_{1-\beta(v_{I(T)B}, v_{I(T)A})} = \frac{F_{\alpha(v_{I(T)B}, v_{I(T)A})}}{\phi^2}$$

$$F_{\beta(v_{I(T)A}, v_{I(T)B})} = 1/F_{1-\beta(v_{I(T)B}, v_{I(T)A})}$$

and find from an F-table the probability  $\beta$  that  $F \geq F_{\beta(v_{I(T)A}, v_{I(T)B})}$

$F_{\alpha(v_{I(T)B}, v_{I(T)A})}$  is the value of the F-distribution with  $v_{I(T)B}$  degrees of freedom associated with the numerator and  $v_{I(T)A}$  degrees of freedom associated with the denominator;  $\alpha$  represents the portion of the F-distribution to the right of the given F-value and

$$v_{I(T)A} = p_A - 1$$

$$v_{I(T)B} = p_B - 1$$

$F_{\beta(v_{I(T)A}, v_{I(T)B})}$  is the value of the F-distribution with  $v_{I(T)A}$  degrees of freedom associated with the numerator and  $v_{I(T)B}$  degrees of freedom associated with the denominator;  $\beta$  represents the portion of the F-distribution to the right of the given F-value.

## b) Repeatabilities of both methods are not equal or $n_A \neq n_B$

Proceed as described in Section 4.9.2.1.2b.

If  $F_{I(T)} \leq F_{\alpha(v_{I(T)B}, v_{I(T)A})}$ , to evaluate  $\beta$ , calculate:

$$F_{1-\beta(v_{I(T)B}, v_{I(T)A})} = \frac{F_{\alpha(v_{I(T)B}, v_{I(T)A})}}{\phi^2}$$

$$F_{\beta(v_{I(T)A}, v_{I(T)B})} = 1/F_{1-\beta(v_{I(T)B}, v_{I(T)A})}$$

and find from an F-table the probability  $\beta$  that  $F \geq F_{\beta(v_{I(T)A}, v_{I(T)B})}$

$F_{\alpha(v_{I(T)B}, v_{I(T)A})}$  is the value of the F-distribution with  $v_{I(T)B}$  degrees of freedom associated with the numerator and  $v_{I(T)A}$  degrees of freedom associated with the denominator,  $\alpha$  represents the portion of the F-distribution to the right of the given F-value and  $v_{I(T)A}$  and  $v_{I(T)B}$  are obtained from eq.(8).

$F_{\beta(v_{I(T)A}, v_{I(T)B})}$  is the value of the F-distribution with  $v_{I(T)A}$  degrees of freedom associated with the numerator and  $v_{I(T)B}$  degrees of freedom associated with the denominator,  $\beta$  represents the portion of the F-distribution to the right of the given F-value and  $v_{I(T)A}$  and  $v_{I(T)B}$  are obtained from eq.(8).

### 4.9.3. Comparison of trueness

#### 4.9.3.1. Based on the minimal number of measurements required

##### 4.9.3.1.1. Comparison of the mean with the certified value of a Reference Material (RM).

The grand mean of each method can be compared with the certified value of the RM used as one of the test samples.

If the uncertainty in the certified value is not taken into account the following tests may be used:

### a) Point hypothesis testing

$$\text{If } |\mu_0 - \bar{\bar{y}}_B| > t_{\alpha/2; p_B - 1} \sqrt{s_{\bar{\bar{y}}_B}^2 / p_B}$$

the difference between the grand mean of the results of the method and the certified value is statistically significant;

$$\text{if } |\mu_0 - \bar{\bar{y}}_B| \leq t_{\alpha/2; p_B - 1} \sqrt{s_{\bar{\bar{y}}_B}^2 / p_B} \quad (9)$$

the difference between the grand mean of the results of the method and the certified value is statistically insignificant ( $\mu_B = \mu_0$ ).

### b) Interval hypothesis testing

Calculate the 90% confidence interval around  $\bar{\bar{y}}_B - \mu_0$ :

$$(\bar{\bar{y}}_B - \mu_0) - t_{0.05; p_B - 1} s_{\bar{\bar{y}}_B} \leq \mu_B - \mu_0 \leq (\bar{\bar{y}}_B - \mu_0) + t_{0.05; p_B - 1} s_{\bar{\bar{y}}_B}$$

$$\text{with } s_{\bar{\bar{y}}_B} = \sqrt{s_{\bar{\bar{y}}_B}^2 / p_B}$$

If this interval is completely included in the acceptance interval  $[-\lambda, \lambda]$ , the difference between the grand mean of the method and the certified value is considered acceptable at the 95% confidence level.

If this interval is not completely included in the acceptance interval  $[-\lambda, \lambda]$ , the difference between the grand mean of the method and the certified value is considered unacceptable at the 95% confidence level.

#### Example:

$$* \mu_0 = 0.6 \quad \bar{\bar{y}}_B = 0.7 \quad s_{\bar{\bar{y}}_B} = 0.1 \quad p_B = 6 \quad \lambda = 0.3$$

$$* v = 6 - 1 = 5 \quad \alpha = 0.05 \quad t_{0.05; 5} = 2.015$$

$$0.1 - 2.015 \times 0.1 \leq \mu_B - \mu_0 \leq 0.1 + 2.015 \times 0.1$$

$$- 0.1015 \leq \mu_B - \mu_0 \leq 0.3015$$

\* Since this interval is not completely included in the acceptance interval  $[- 0.3, 0.3]$  the difference between the grand mean of the method and the certified value is considered unacceptable at the 95% confidence level.

#### 4.9.3.1.2. Comparison between the means of method A and B

##### a) Point hypothesis testing

$$\text{If } \frac{|\bar{y}_A - \bar{y}_B|}{s_d} > t_{0.025;v_d} \quad (10)$$

the difference between the means of method A and method B is statistically significant at  $\alpha = 0.05$ .

$$\text{If } \frac{|\bar{y}_A - \bar{y}_B|}{s_d} \leq t_{0.025;v_d} \quad (11)$$

the difference between the mean of method A and method B is statistically insignificant at  $\alpha = 0.05$  ( $\mu_A = \mu_B$ ).

With  $t_{0.025;v_d}$  is the one-sided tabulated t-value at significance level 0.025 and degrees of freedom  $v_d$ .

The computation of  $s_d$  as well as the number of degrees of freedom  $v_d$  associated with  $s_d$ , depends on whether or not the variance of the day means for both methods are equal ( $\sigma_{\bar{y}_A}^2 = \sigma_{\bar{y}_B}^2$ ). This is evaluated as follows:

$$F = \frac{s_1^2}{s_2^2} \quad (12)$$

with  $s_1^2$  the largest of  $s_{\bar{y}_A}^2$  and  $s_{\bar{y}_B}^2$

Compare F with  $F_{\alpha/2(v_1, v_2)}$  where

$$\begin{aligned} v_1 &= p_A - 1 \quad \text{if } s_{\bar{y}_A}^2 > s_{\bar{y}_B}^2 \\ &= p_B - 1 \quad \text{if } s_{\bar{y}_B}^2 > s_{\bar{y}_A}^2 \end{aligned}$$

$$\begin{aligned} v_2 &= p_B - 1 \quad \text{if } s_{\bar{y}_A}^2 > s_{\bar{y}_B}^2 \\ &= p_A - 1 \quad \text{if } s_{\bar{y}_B}^2 > s_{\bar{y}_A}^2 \end{aligned}$$

If  $F \leq F_{\alpha/2}(v_1, v_2)$  there is no evidence that the variance of the day means of both methods is different at  $\alpha=0.05$ .

In that case  $s_d$  in eq. (10) is obtained as follows:

$$s_d = \sqrt{s_p^2 \left( \frac{1}{p_A} + \frac{1}{p_B} \right)} \quad (13)$$

with

$$s_p^2 = \frac{(p_A - 1)s_{\bar{y}_A}^2 + (p_B - 1)s_{\bar{y}_B}^2}{p_A + p_B - 2} \quad (14)$$

and the number of degrees of freedom associated with  $s_d$  is  $v_d = p_A + p_B - 2$

If  $F > F_{\alpha/2}(v_1, v_2)$  there is evidence that the variance of the day means of both methods is different at  $\alpha=0.05$ .

In that case  $s_d$  in eq. (10) is obtained as follows:

$$s_d = \sqrt{\frac{s_{\bar{y}_A}^2}{p_A} + \frac{s_{\bar{y}_B}^2}{p_B}} \quad (15)$$

and the number of degrees of freedom associated with  $s_d$  is then calculated by applying the Satterthwaite approximation:

$$v_d = \frac{\left( s_d^2 \right)^2}{\left( \frac{s_{\bar{y}_A}^2}{p_A} \right)^2 / (p_A - 1) + \left( \frac{s_{\bar{y}_B}^2}{p_B} \right)^2 / (p_B - 1)} \quad (16)$$

## b) Interval hypothesis testing

Calculate the 90% confidence interval around  $(\bar{\bar{y}}_A - \bar{\bar{y}}_B)$ :

$$(\bar{\bar{y}}_A - \bar{\bar{y}}_B) - t_{0.05; v_d} s_d \leq \mu_A - \mu_B \leq (\bar{\bar{y}}_A - \bar{\bar{y}}_B) + t_{0.05; v_d} s_d$$

With  $s_d$  calculated according to eq.(13) or eq. (15) depending on whether the variance of the day means for both methods are equal or different, respectively.

If this interval is completely included in the acceptance interval  $[-\lambda, \lambda]$ , the difference between the grand means of method A and method B is considered

acceptable at the 95% confidence level. If this interval is not completely included in the acceptance interval  $[-\lambda, \lambda]$ , the difference between the grand means of method A and method B is not considered acceptable at the 95% confidence level.

Example:

$$* \bar{y}_A = 0.6 \quad \bar{y}_B = 0.7 \quad s_{\bar{y}_A} = 0.175 \quad s_{\bar{y}_B} = 0.175$$

$$p_A = 6 \quad p_B = 6 \quad \lambda = 0.3$$

Since the variance of the day means of both methods can be considered to be equal  $s_d$  is obtained from eq. (13):

$$* s_p^2 = 0.0306 \quad s_d = 0.101$$

$$v_d = 10 \quad t_{0.05;10} = 1.812$$

$$-0.1 - 1.812 \times 0.101 \leq \mu_A - \mu_B \leq -0.1 + 1.812 \times 0.101$$

$$-0.283 \leq \mu_A - \mu_B \leq 0.083$$

\* Since this interval is completely included in the acceptance interval  $[-0.3, 0.3]$  the difference between the grand means of method A and method B is considered acceptable at the 95% confidence level.

#### 4.9.3.2. Based on a user-defined number of measurements

##### 4.9.3.2.1. Comparison of the mean with the certified value of a Reference Material (RM)

###### a) Point hypothesis testing

The trueness (bias) is evaluated as described in Section 4.9.3.1.1a but additionally the probability  $\beta$  of not detecting a bias which in reality is equal to  $\lambda$  is computed.

$$\text{Therefore if } |\mu_0 - \bar{y}_B| \leq t_{\alpha/2; p_B-1} \sqrt{s_{\bar{y}_B}^2 / p_B} \quad (\text{see eq.(9)})$$

calculate:

\* the upper limit that leads to the acceptance that  $\mu_B = \mu_0$

$$UL = t_{\alpha/2; p_B - 1} s_{\bar{y}_B} \quad s_{\bar{y}_B} = \sqrt{s_{y_B}^2 / p_B}$$

\* for the distribution centered around  $\lambda$  find the probability to obtain a value smaller than UL. Therefore calculate

$$t_{\beta} = \frac{|\lambda - UL|}{s_{\bar{y}_B}}$$

and from the t-distribution with  $v = p_B - 1$  find the probability that  $t > t_{\beta}$  if  $\lambda - UL > 0$  and find the probability that  $t < t_{\beta}$  if  $\lambda - UL < 0$

**Example:**

\*  $\mu_0 = 0.6 \quad \bar{y}_B = 0.7 \quad s_{\bar{y}_B} = 0.1 \quad p_B = 6 \quad \lambda = 0.3$

\*  $v = 6 - 1 = 5 \quad \alpha = 0.05 \quad \rightarrow \quad t_{0.025; 5} = 2.571$

$$|0.6 - 0.7| < 2.571 \times 0.1$$

$$0.1 < 0.2571$$

the difference is not significant

\* the probability that a real difference  $\lambda = 0.3$  would not be detected is obtained as follows:

$$UL = 0.2571$$

$$t_{\beta} = \frac{|0.3 - 0.2571|}{0.1} = 0.429$$

Since  $\lambda - UL > 0$  the probability that  $t > t_{\beta}$  is 34%.

Therefore the probability of not detecting a bias equal to 0.3 if this is real is 34%.

**b) Interval hypothesis testing**

Proceed as described in Section 4.9.3.1.1b.

**4.9.3.2.2. Comparison of the means of method A and B.**

**a) Point hypothesis testing**

The trueness (bias) is calculated as described in Section 4.9.3.1.2a but additionally the probability  $\beta$  of not detecting a bias which in reality is equal to  $\lambda$  is computed.

Therefore if  $|\bar{y}_A - \bar{y}_B| / s_d \leq t_{0.025;v_d}$  (see eq. (11)) calculate

\* the upper limit that leads to the acceptance that  $\mu_A = \mu_B$ :

$$UL = t_{\alpha/2;v_d} s_d$$

\* for the distribution centered around  $\lambda$  find the probability to obtain a value smaller than UL. Therefore calculate

$$t_\beta = \frac{|\lambda - UL|}{s_d}$$

and from the t-distribution with  $v_d$  degrees of freedom find the probability that

$t > t_\beta$  if  $\lambda - UL > 0$  and find the probability that  $t < t_\beta$  if  $\lambda - UL < 0$

With  $s_d$  calculated according to eq. (13) or eq. (15), depending on whether the variance of the day means for both methods are equal or different, respectively.

Example :

$$* \bar{y}_A = 0.6 \quad \bar{y}_B = 0.7 \quad s_{\bar{y}_A} = 0.175 \quad s_{\bar{y}_B} = 0.175$$

$$p_A = 6 \quad p_B = 6 \quad \lambda = 0.3$$

Since the variance of the day means of both methods can be considered to be equal  $s_d$  is obtained from eq.(13).

$$* s_p^2 = 0.0306 \quad s_d = 0.101$$

$$v_d = 10 \quad t_{0.025;10} = 2.228$$

\* The means for methods A and B are not significantly different since (see eq. (11)):

$$\frac{|0.6 - 0.7|}{0.101} = 0.991 < 2.228$$

\* the probability that a real difference  $\lambda = 0.3$  would not be detected is obtained as follows :

$$UL = 2.228 \times 0.101 = 0.225$$

$$t_\beta = \frac{|0.3 - 0.225|}{0.101} = 0.743$$

$$\rightarrow \beta = 24\%$$

## **b) Interval hypothesis testing**

Proceed as in Section 4.9.3.1.2b

## **5. Examples**

Two examples will illustrate the approach discussed. In the first example the minimal number of measurements to be performed in the method comparison is determined. The second example starts from a user-defined number of measurements and the probability  $\beta$  to adopt an unacceptable method is then evaluated.

### **5.1. Example 1**

#### **5.1.1. Background**

##### **5.1.1.1. Measurement methods**

The example is fictitious.

Method A is a Karl Fischer method, method B a vacuum oven method for the determination of moisture in cheese.

A laboratory uses method A but developed method B as an alternative. The laboratory wants to compare the performance of both methods. The results are expressed as % moisture.

##### **5.1.1.2. Experimental design**

The material is a cheese, analyzed with both methods. Each day during  $p_A$  days, 2 independent samples ( $n_A = 2$ ) from the cheese will be analyzed with method A. Each day during  $p_B$  days, 2 independent samples ( $n_B = 2$ ) from the cheese will be analyzed with method B.

It is decided to take  $p_A = p_B$ .

##### **5.1.2. Requirements**

$$\lambda = 0.50\%$$

$$\rho = \phi = 3$$

### 5.1.3. Determination of $p_A (= p_B)$

For method A an estimate of the precision ( $s_r^2$  and  $s_t^2$ ) is available:

$$s_r^2 = 0.023 \quad s_t^2 = 0.08$$

The minimum number of days required:

#### *a) For the trueness requirement*

$$0.5 = (t_{\alpha/2} + t_{\beta}) \sqrt{2(0.08 + 0.023/2)/p_A}$$

With  $p_A = 6$ ,  $(t_{\alpha/2} + t_{\beta}) = (2.228 + 0.879) = 3.107$  and  $\lambda = 0.543$ ; with  $p_A = 7$ ,

$(t_{\alpha/2} + t_{\beta}) = 3.051$  and  $\lambda = 0.493$ . Hence  $p_A = p_B = 7$ .

*NOTE: the use of a constant multiplication factor equal to 3 would yield  $p_A = p_B = 7$ .*

#### *b) For the precision requirement*

From Table 1, it can be seen that  $\rho = 3$  or  $\phi = 3$  is given by  $v_A = v_B = 6$ .

To compare repeatability standard deviations:

$$v_A = p_A \text{ and } v_B = p_B, \text{ so } p_A = p_B = 6$$

To compare between-day mean squares:

$$v_A = p_A - 1 \text{ and } v_B = p_B - 1, \text{ so } p_A = p_B = 7$$

#### *c) Conclusion*

The minimum number of days required (with two measurements per day) is 7.

### 5.1.4. The data

The data are summarized in Table 4.

**Table 4**

Test results (Example 1)

Day	Method A		Method B	
	$y_{ij}$	$\bar{y}_i$	$y_{ij}$	$\bar{y}_i$
1	39.68	39.725	39.29	39.325
	39.77		39.36	
2	39.08	39.230	39.51	39.445
	39.38		39.38	
3	40.39	40.360	39.45	39.470
	40.33		39.49	
4	39.92	40.060	39.29	39.325
	40.20		39.36	
5	40.34	40.115	39.83	39.855*
	39.89		39.88	
6	40.12	40.190	39.44	39.445
	40.26		39.45	
7	39.43	39.485	39.45	39.490
	39.54		39.53	

$$\bar{\bar{y}}_A = 39.881$$

$$\bar{\bar{y}}_B = 39.479$$

### 5.1.5. Graphical presentation

A graphical presentation of the data from Table 4 for methods A and B is given in Figures 1 and 2 respectively. Figure 3 represents for both methods the absolute difference between the duplicates and Figure 4 the 7 day means. (All figures are presented in the Appendix II.)

Figures 1 and 2 do not invite specific remarks.

Inspection of Figure 3 reveals that the repeatability for method B is at least as good as that for method A since the difference between the duplicates for method B are not larger than those of method A.

From Figure 4 it follows that the mean of day 5 for method B might be outlying. However Figure 4 also indicates that the between-day precision for method B is at least as good as that for method A since the spread of day means around the grand mean for the former method is less than the spread for the latter method.

Nevertheless, to illustrate the calculations, the statistical analysis for the comparison of the precision of both methods will be carried out.

### 5.1.6. Investigation of outliers

Grubbs' tests were applied to the day means.

No single or double stragglers or outliers were found for method A. For method B the single Grubbs' test applied on the mean of day 5 is significant at the 5% level but not at the 1% level. Indeed

$$G = \frac{39.855 - 39.479}{0.1786} = 2.105$$

which is to be compared with Grubbs' critical values for  $p = 7$  at 5% (2.020) and 1% (2.139).

Therefore since this observation is considered as a straggler it is retained but indicated with an asterisk in Table 4.

### 5.1.7. Calculation of the variances

Tables 5 and 6 summarize the calculation of the variances for methods A and B, respectively.

**Table 5**  
Calculation of the variances for method A

ANOVA table

Source	Mean squares	Estimate of
Day	$MS_D = 0.3389$	$\sigma_{rA}^2 + n_A \sigma_{tA}^2$
Residual	$MS_E = 0.0296$	$\sigma_{rA}^2$

#### Calculation of variances

- The repeatability variance

$$s_{rA}^2 = 0.0296 \qquad df = 7(2-1) = 7$$

- Variance component between days (between-day variance)

$$s_{tA}^2 = \frac{0.3389 - 0.0296}{2} = 0.1546$$

- Time-different intermediate precision

$$s_{I(T)A}^2 = s_{tA}^2 + s_{rA}^2 = 0.1842$$

- Variance of the means  $\bar{y}_i$

$$s_{\bar{y}A}^2 = s_{tA}^2 + s_{rA}^2 / n_A = 0.1694 \qquad df = (7-1) = 6$$

**Table 6**  
Calculation of the variances for method B

ANOVA table

Source	Mean squares	Estimate of
Day	$MS_D = 0.0638$	$\sigma_{rB}^2 + n_B \sigma_{tB}^2$
Residual	$MS_E = 0.0027$	$\sigma_{rB}^2$

**Calculation of variances**

- The repeatability variance

$$s_{rB}^2 = 0.0027$$

$$df = 7(2-1) = 7$$

- Variance component between days (between-day variance)

$$s_{tB}^2 = \frac{0.0638 - 0.0027}{2} = 0.0306$$

- Time-different intermediate precision

$$s_{I(T)B}^2 = s_{tB}^2 + s_{rB}^2 = 0.0333$$

- Variance of the means  $\bar{y}_i$

$$s_{\bar{y}B}^2 = s_{tB}^2 + s_{rB}^2/n_B = 0.0319$$

$$df = (7-1) = 6$$

### 5.1.8. Comparison of precision

a) Repeatability

$$F_r = \frac{0.0027}{0.0296} = 0.091$$

This is to be compared with  $F_{0.05(7,7)} = 3.79$ . Since  $F_r < 3.79$  there is no evidence that the repeatability of method B is worse than that of method A.

b) Time-different intermediate precision

- Check whether  $\sigma_{rA}^2 = \sigma_{rB}^2$

$$F = \frac{0.0296}{0.0027} = 10.96$$

This is to be compared with  $F_{0.025(7,7)} = 4.99$ .

Since  $F > 4.99$  there is evidence that the repeatabilities of both methods are different (in fact the repeatability for method B is better than for method A).

- The repeatabilities of both methods being different a comparison of the time-different intermediate precision is performed as follows:

$$F_{I(T)} = \frac{s_{I(T)B}^2}{s_{I(T)A}^2} = \frac{0.0333}{0.1842} = 0.181$$

$$v_{I(T)A} = \frac{0.1842^2}{(0.3389/2)^2/6 + (0.0296/2)^2/7} = 7$$

$$v_{I(T)B} = \frac{0.0333^2}{(0.0638/2)^2/6 + (0.0027/2)^2/7} = 6$$

$F_{I(T)}$  is to be compared with  $F_{0.05(6,7)} = 3.87$ . Since  $F_{I(T)} < 3.87$  there is no evidence that the time-different intermediate precision of method B is worse than that of method A (in fact the time-different intermediate precision for method B is better than for method A).

### 5.1.9. Comparison of trueness (bias)

This is done by comparing the means of method A and B.

- Check whether  $\sigma_{\bar{y}_A}^2 = \sigma_{\bar{y}_B}^2$

$$F = \frac{0.1694}{0.0319} = 5.31$$

This is to be compared with  $F_{0.025(6,6)} = 5.82$ . Since  $F < 5.82$ , there is no evidence that the variances of the day means obtained with the two methods are different.

- Therefore the variances can be pooled (Eq.(14)) and  $s_d$  is obtained from eq. (13):

$$s_p^2 = \frac{(6 * 0.1694) + (6 * 0.0319)}{12} = 0.1007$$

$$s_d = \sqrt{0.1007 \left( \frac{1}{7} + \frac{1}{7} \right)} = 0.1696$$

a) Point hypothesis testing

$$\frac{|\bar{\bar{y}}_A - \bar{\bar{y}}_B|}{s_d} = \frac{|39.881 - 39.479|}{0.1696} = 2.37$$

This is to be compared with  $t_{0.025;12} = 2.18$ . Since  $2.37 > 2.18$ , the difference between the means of the two methods is statistically significant at  $\alpha = 0.05$ .

b) Interval hypothesis testing

Calculate the 90% confidence interval around  $|\bar{\bar{y}}_A - \bar{\bar{y}}_B|$ :

$$(39.881 - 39.479) - 0.1696 * t_{0.05;12} \leq \mu_A - \mu_B \leq (39.881 - 39.479) + 0.1696 * t_{0.05;12}$$

$$0.402 - (0.1696 * 1.782) \leq \mu_A - \mu_B \leq 0.402 + (0.1696 * 1.782)$$

$$0.100 \leq \mu_A - \mu_B \leq 0.704$$

Since this interval is not completely included in the acceptance interval  $[-0.5, 0.5]$ , the difference between the grand means of method A and method B is considered unacceptable at the 95% confidence level.

*NOTE:*

- Consider the hypothetical case that  $|\bar{y}_A - \bar{y}_B| = 0.30$  and  $s_d = 0.10$  (which is smaller than the value obtained from the variance estimates used in the determination of the number of measurements in eq. (3)). The point hypothesis testing approach would yield the same conclusion as above while from the interval hypothesis testing with the 90% confidence interval around  $|\bar{y}_A - \bar{y}_B|$ :

$$\begin{aligned} 0.30 - (0.10 * 1.782) &\leq \mu_A - \mu_B \leq 0.30 + (0.10 * 1.782) \\ 0.122 &\leq \mu_A - \mu_B \leq 0.478 \end{aligned}$$

the conclusion would be that the difference between the biases of the two methods is acceptable because the interval is completely included in  $[-0.5, 0.5]$ .

- Consider the hypothetical case that  $|\bar{y}_A - \bar{y}_B| = 0.30$  and  $s_d = 0.20$  (which is larger than the value obtained from the variance estimates used in the determination of the number of measurements in eq. (3)).

Point hypothesis testing:

$$\frac{|\bar{y}_A - \bar{y}_B|}{s_d} = \frac{0.30}{0.20} = 1.50$$

indicating (since  $1.50 < 2.18$ ) that the difference between the means of method A and B is statistically not significant.

The interval hypothesis testing with the 90% confidence interval around  $|\bar{y}_A - \bar{y}_B|$ :

$$\begin{aligned} 0.30 - (0.20 * 1.782) &\leq \mu_A - \mu_B \leq 0.30 + (0.20 * 1.782) \\ -0.056 &\leq \mu_A - \mu_B \leq 0.656 \end{aligned}$$

indicating that the difference between the biases of both methods is unacceptable because the interval is not completely included in  $[-0.5, 0.5]$ .

- The differences with both approaches in these two examples are due to the fact that the experimentally obtained  $s_d$  ( $s_d = 0.10$  and  $0.20$  for the first and second examples, respectively) does not correspond with the value obtained ( $s_d = 0.16$ ) from the variance estimates used in eq. (3) for the determination of the minimal number of measurements. Therefore it seems that despite the fact that the minimal number of measurements, required to control the  $\beta$ -error, have been used in the experiments an evaluation of the results by means of the interval hypothesis testing is to be preferred.

## **5.2. Example 2**

### **5.2.1. Background**

#### **5.2.1.1. Measurement methods**

Method A is a flame AAS method, method B a CZE method for the determination of Ca in total diet.

A laboratory uses method A but developed method B as an alternative. The laboratory wants to compare the performance of both methods.

Results are expressed as mg/100 g dry weight.

#### **5.2.1.2. Experimental design**

The material is a total diet, analyzed with both methods.

Each day, during 6 days ( $p_A = 6$ ), two independent samples ( $n_A = 2$ ) from the total diet are analyzed with method A.

Each day, during 7 days ( $p_B = 7$ ), two independent samples ( $n_B = 2$ ) from the total diet are analyzed with method B.

### **5.2.2. The data**

The data obtained are summarized in Table 7.

### **5.2.3. Graphical presentation**

A graphical presentation of the data from Table 7 for methods A and B is given in Figures 5 and 6 respectively. Figure 7 represents for both methods the absolute difference between the duplicates and Figure 8 the day means. (All figures are presented in the Appendix II.)

The figures do not invite specific remarks.

### **5.2.4. Investigation of outliers**

Grubbs' tests were applied to the day means.

No single or double stragglers or outliers were found for both method A and method B.

### **5.2.5. Calculation of the variances**

Tables 8 and 9 summarize the calculation of the variances for methods A and B, respectively.

**Table 7**

Test results (Example 2)

Day	Method A		Method B	
	$y_{ij}$	$\bar{y}_i$	$y_{ij}$	$\bar{y}_i$
1	195	197	186	190
	199		194	
2	199	205.5	185	178.5
	212		172	
3	206	212	180	182
	218		184	
4	187	190	188	197.5
	193		207	
5	206	202	220	207
	198		194	
6	196	198	197	207
	200		217	
7			188	190.5
			193	
	$\bar{\bar{y}}_A = 200.75$		$\bar{\bar{y}}_B = 193.21$	

**Table 8**

Calculation of the variances for method A

ANOVA table

Source	Mean squares	Estimate of
Day	$MS_D = 115.15$	$\sigma_{rA}^2 + n_A \sigma_{tA}^2$
Residual	$MS_D = 37.08$	$\sigma_{rA}^2$

**Calculation of variances**

- The repeatability variance

$$s_{rA}^2 = 37.08$$

$$df = 6(2-1) = 6$$

- Variance component between days (between-day variance)

$$s_{tA}^2 = \frac{115.15 - 37.08}{2} = 39.035$$

- Time-different intermediate precision

$$s_{I(T)A}^2 = s_{rA}^2 + s_{tA}^2 = 76.115$$

- Variance of the means  $\bar{y}_i$ 

$$s_{\bar{y}A}^2 = s_{tA}^2 + s_{rA}^2 / n_A = 57.575$$

$$df = (6-1) = 5$$

**Table 9**  
Calculation of the variances for method B

ANOVA table

Source	Mean squares	Estimate of
Day	$MS_D = 252.81$	$\sigma_{rB}^2 + n_B \sigma_{tB}^2$
Residual	$MS_E = 122.21$	$\sigma_{rB}^2$

**Calculation of variances**

- The repeatability variance

$$s_{rB}^2 = 122.21$$

$$df = 7(2-1) = 7$$

- Variance component between days (between-day variance)

$$s_{tB}^2 = \frac{252.81 - 122.21}{2} = 65.30$$

- Time-different intermediate precision

$$s_{I(T)B}^2 = s_{rB}^2 + s_{tB}^2 = 187.51$$

- Variance of the mean  $\bar{y}_i$

$$s_{\bar{y}B}^2 = s_{tB}^2 + s_{rB}^2 / n_B = 126.405$$

$$df = (7-1) = 6$$

### 5.2.6. Comparison of precision

a) Repeatability

$$F_r = \frac{122.21}{37.08} = 3.30$$

This is to be compared with  $F_{0.05(7,6)} = 4.21$ . Since  $F_r < 4.21$  there is no evidence that the repeatability of the CZE method is worse than the repeatability of the AAS method.

Suppose that the laboratory considers a ratio  $\rho^2 = \sigma_{rB}^2 / \sigma_{rA}^2 = 4$  to be important. The probability  $\beta$  that, if in reality  $\sigma_{rB}^2 = 4\sigma_{rA}^2$  ( $\sigma_{rB} = 2\sigma_{rA}$ ), the test will lead to the conclusion that the repeatabilities are not significantly different is obtained from:

$$F_{1-\beta(7,6)} = \frac{F_{0.05(7,6)}}{\rho^2} = \frac{4.21}{4} = 1.0525$$

$$F_{\beta(6,7)} = 1/F_{1-\beta(7,6)} = 1/1.0525 = 0.95$$

From the F-distribution  $\beta$  is found to be 52%.

This means that if in reality  $\sigma_{rB}^2 = 4\sigma_{rA}^2$ , the probability that, with the used experimental set-up, the laboratory will conclude that both methods have similar repeatabilities is 52%.

b) Time-different intermediate precision

- Check whether  $\sigma_{rA}^2 = \sigma_{rB}^2$

$$F = \frac{122.21}{37.08} = 3.30$$

This is to be compared with  $F_{0.025(7,6)} = 5.70$ . Since  $F < 5.70$  there is no evidence that the repeatabilities of both methods are different.

- The repeatabilities of both methods not being different a comparison of the time-different intermediate precision is performed as follows:

$$F_{I(T)} = \frac{s_{\bar{y}_B}^2}{s_{\bar{y}_A}^2} = \frac{126.405}{57.575} = 2.20$$

This is to be compared with  $F_{0.05(6,5)} = 4.95$ . Since  $F_{I(T)} < 4.95$  there is no evidence that the time-different intermediate precision of the CZE method is worse than that of the AAS method.

Suppose that the laboratory considers a ratio  $\phi^2 = \sigma_{\bar{y}_B}^2 / \sigma_{\bar{y}_A}^2 = 4$  to be important. The probability  $\beta$  that, if in reality  $\sigma_{\bar{y}_B}^2 = 4\sigma_{\bar{y}_A}^2$  or ( $\sigma_{\bar{y}_B} = 2\sigma_{\bar{y}_A}$ ), the test will lead to the conclusion that the time-different intermediate precisions are not significantly different is obtained from:

$$F_{1-\beta(6,5)} = \frac{F_{0.05(6,5)}}{\phi^2} = \frac{4.95}{4} = 1.2375$$

$$F_{\beta(5,6)} = 1/F_{1-\beta(6,5)} = 0.808$$

From the F-distribution  $\beta$  is found to be 58%.

### 5.2.7. Comparison of trueness (bias)

This is done by comparing the means of method A and B.

- Check whether  $\sigma_{\bar{y}_A}^2 = \sigma_{\bar{y}_B}^2$

$$F = \frac{126.405}{57.575} = 2.20$$

This is to be compared with  $F_{0.025(6,5)} = 6.98$ . Since  $F < 6.98$ , there is no evidence that the variances of the day means obtained with the two methods are different.

- Therefore the variances can be pooled (Eq.(14)) and  $s_d$  is obtained from eq. (13):

$$s_p^2 = \frac{5 * 57.575 + 6 * 126.405}{11} = 95.12$$

$$s_d = \sqrt{95.12 \left( \frac{1}{6} + \frac{1}{7} \right)} = 5.43$$

a) Point hypothesis testing

$$\frac{|\bar{y}_A - \bar{y}_B|}{s_d} = \frac{|200.75 - 193.21|}{5.43} = 1.39$$

This is to be compared with  $t_{0.025;11} = 2.20$ . Since  $1.39 < 2.20$ , the difference between the means of the two methods is not statistically significant at  $\alpha = 0.05$ . Suppose that the laboratory considers a difference of 10 ( $\lambda = 10$ ) to be important. The probability  $\beta$  that, if in reality  $\lambda = 10$ , the test will lead to the conclusion that the CZE method is not biased is obtained from:

$$t_\beta = \frac{|\lambda - UL|}{s_d} = \frac{|10 - 11.95|}{5.43} = 0.359$$

$$\text{with } UL = t_{\alpha/2; (p_A + p_B - 2)} s_d = 2.20 \times 5.43 = 11.95$$

Since  $\lambda - UL < 0$ ,  $\beta$  is found from the t-distribution as the probability that  $t < t_\beta$ . Therefore  $\beta = 64\%$ .

b) Interval hypothesis testing

Calculate the 90% confidence interval around  $|\bar{y}_A - \bar{y}_B|$ :

$$\begin{aligned} (\bar{y}_A - \bar{y}_B) - t_{0.05; (p_A + p_B - 2)} s_d &\leq \mu_A - \mu_B \leq (\bar{y}_A - \bar{y}_B) + t_{0.05; (p_A + p_B - 2)} s_d \\ 7.54 - 1.796 \times 5.43 &\leq \mu_A - \mu_B \leq 7.54 + 1.796 \times 5.43 \\ - 2.212 &\leq \mu_A - \mu_B \leq 17.292 \end{aligned}$$

Since this interval is not completely included within the acceptance interval  $[-10, 10]$  we conclude that the difference between the means of both methods is unacceptable. There is a higher than 5% probability that the difference between the means is larger than 10 (or smaller than -10).

## REFERENCES

- [1] C. Hartmann, J. Smeyers-Verbeke, W. Penninckx, Y. Vander Heyden, P. Vankeerberghen, and D. L. Massart, *Anal. Chem.* 67 (1995) 4491.
- [2] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier, Amsterdam, 1997.
- [3] F. E. Satterthwaite, *Biom. Bull.* 2 (1946) 110.
- [4] International Organization for Standardization, *Accuracy (Trueness and Precision) of Measurement methods and results*, ISO/DIS 5725-2, Geneva (1994).

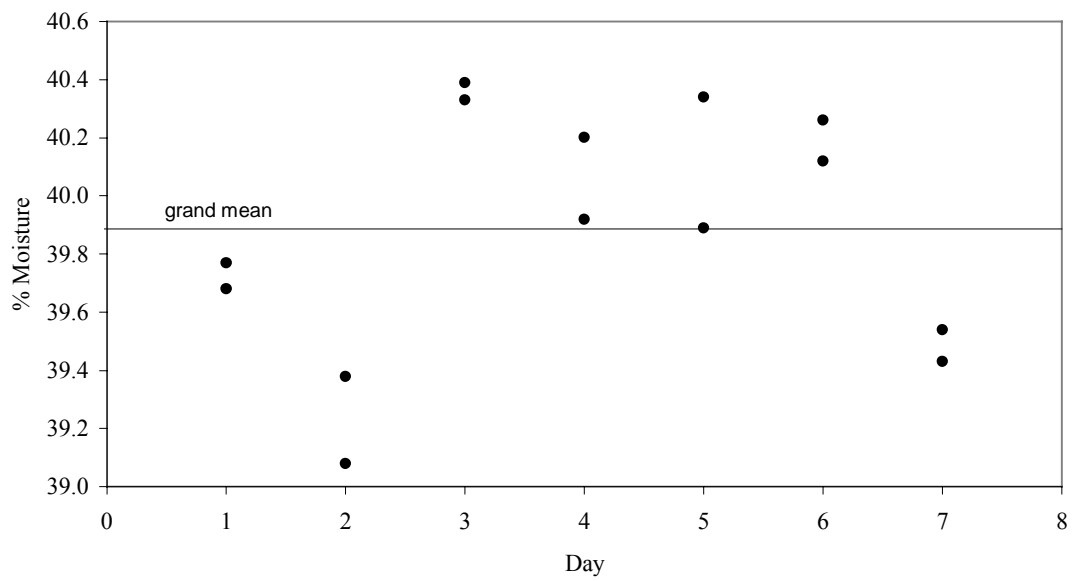
**APPENDIX I**

**Table 10**  
Critical values for Grubbs' test [4]

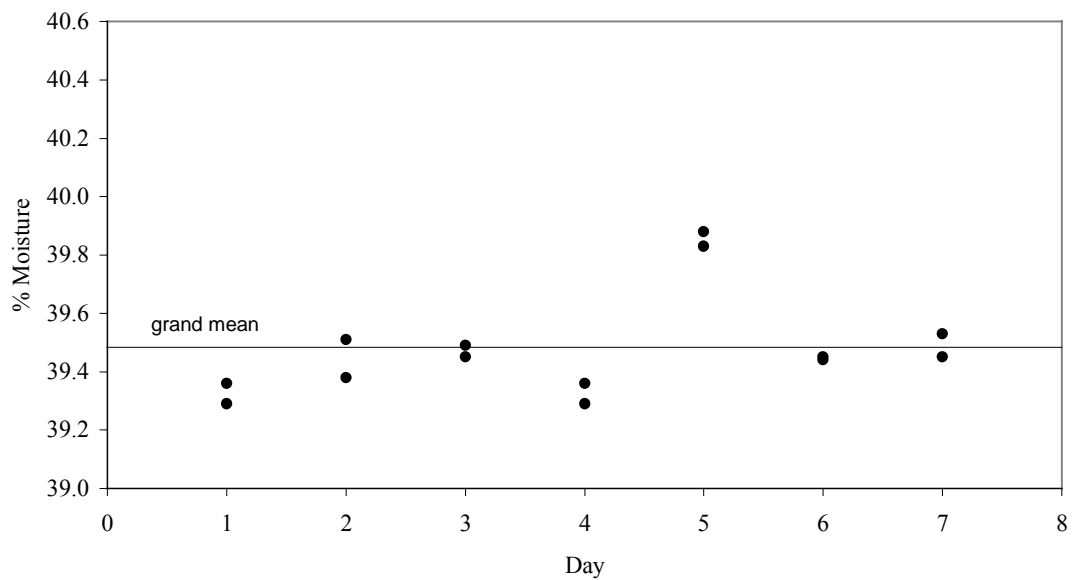
p	One largest or one smallest		Two largest or two smallest	
	Upper 1%	Upper 5%	Lower 1%	Lower 5%
3	1.155	1.155	-	-
4	1.496	1.481	0.0000	0.0002
5	1.764	1.715	0.0018	0.0090
6	1.973	1.887	0.0116	0.0349
7	2.139	2.020	0.0308	0.0708
8	2.274	2.126	0.0563	0.1101
9	2.387	2.215	0.0851	0.1492
10	2.482	2.290	0.1150	0.1864
11	2.564	2.355	0.1448	0.2213
12	2.636	2.412	0.1738	0.2537
13	2.699	2.462	0.2016	0.2836
14	2.755	2.507	0.2280	0.3112
15	2.806	2.549	0.2530	0.3367
16	2.852	2.585	0.2767	0.3603
17	2.894	2.620	0.2990	0.3822
18	2.932	2.651	0.3200	0.4025
19	2.968	2.681	0.3398	0.4214
20	3.001	2.709	0.3585	0.4391
21	3.031	2.733	0.3761	0.4556
22	3.060	2.758	0.3927	0.4711
23	3.087	2.781	0.4085	0.4857
24	3.112	2.802	0.4234	0.4994
25	3.135	2.822	0.4376	0.5123
26	3.157	2.841	0.4510	0.5245
27	3.178	2.859	0.4638	0.5360
28	3.199	2.876	0.4759	0.5470
29	3.218	2.893	0.4875	0.5574
30	3.236	2.908	0.4985	0.5672
31	3.253	2.924	0.5091	0.5766
32	3.270	2.938	0.5192	0.5856
33	3.286	2.952	0.5288	0.5941
34	3.301	2.965	0.5381	0.6023
35	3.316	2.979	0.5469	0.6101
36	3.330	2.991	0.5554	0.6175
37	3.343	3.003	0.5636	0.6247
38	3.356	3.014	0.5714	0.6316
39	3.369	3.025	0.5789	0.6382
40	3.381	3.036	0.5862	0.6445

p = number of days

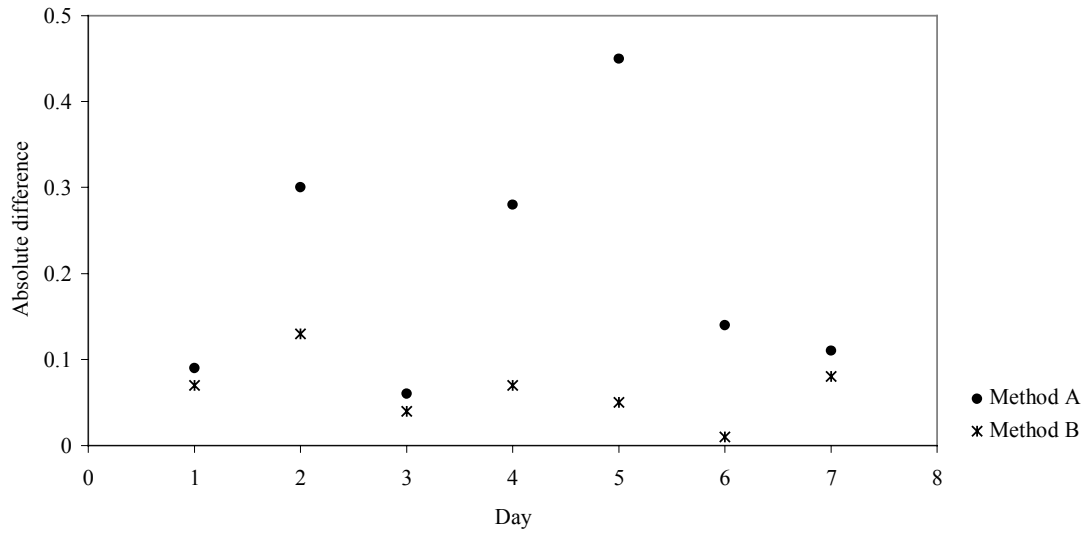
## APPENDIX II



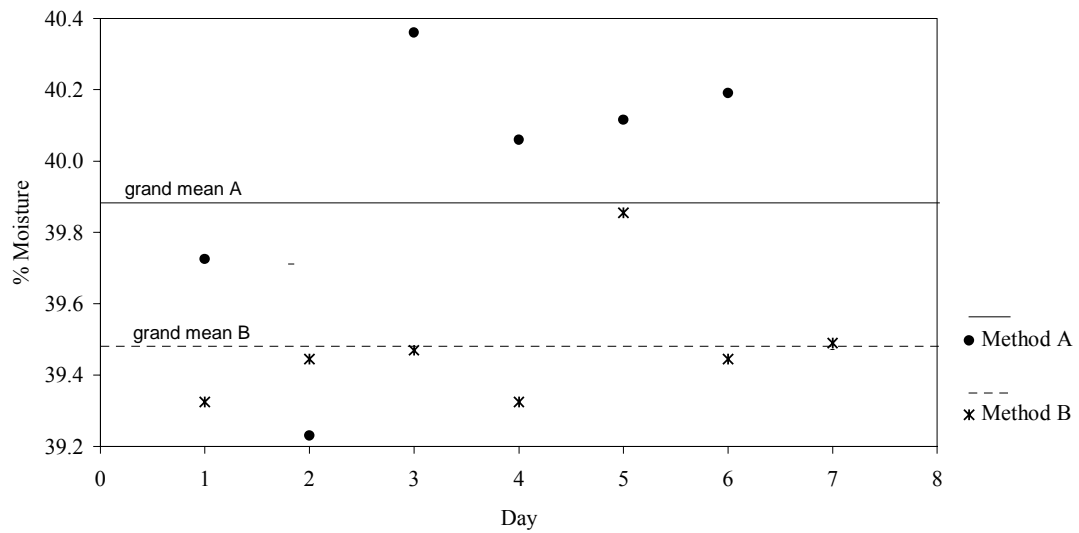
**Figure 1** Results for method A (Example 1)



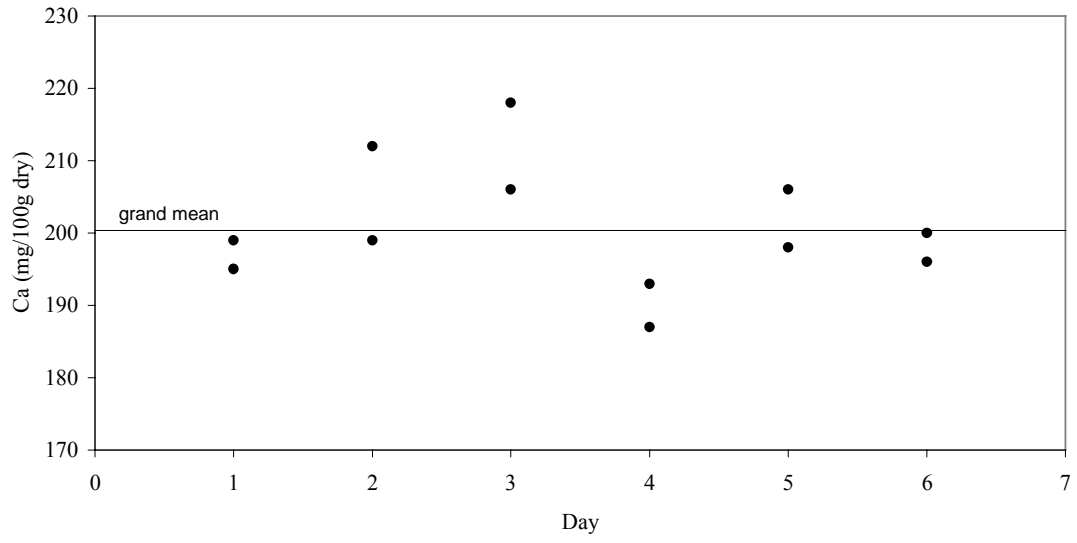
**Figure 2** Results for method B (Example 1)



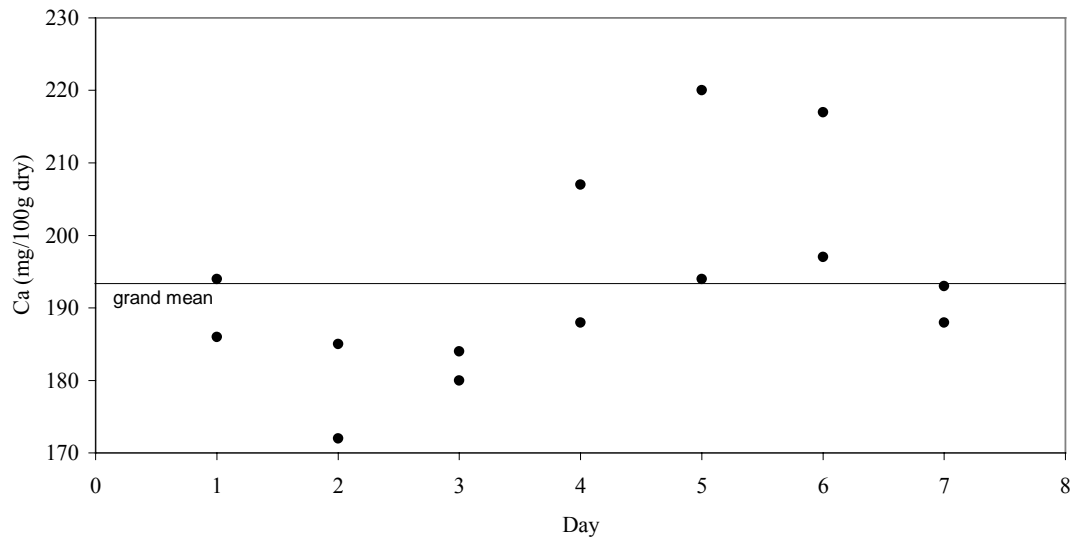
**Figure 3** Absolute differences between duplicates (Example 1)



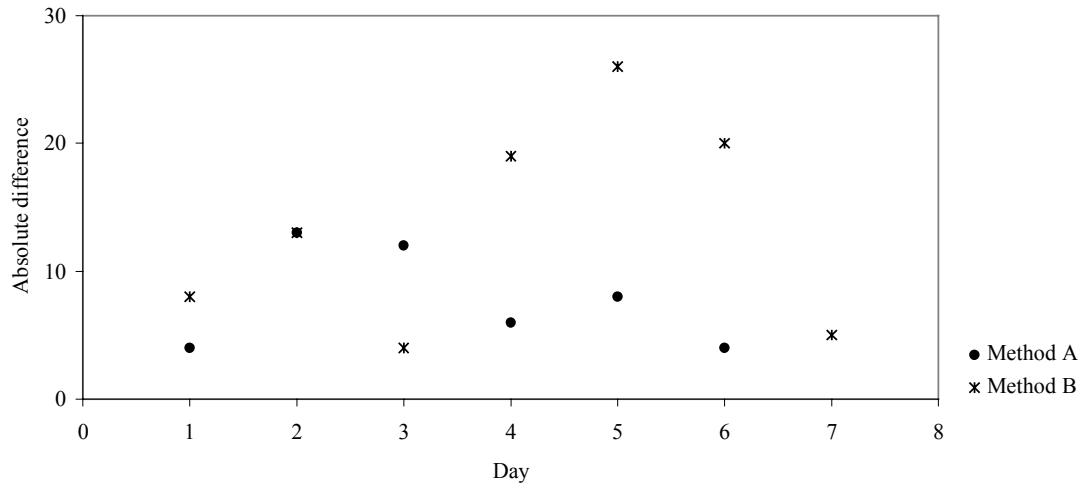
**Figure 4** Day means and the grand means obtained with the two methods (Example 1)



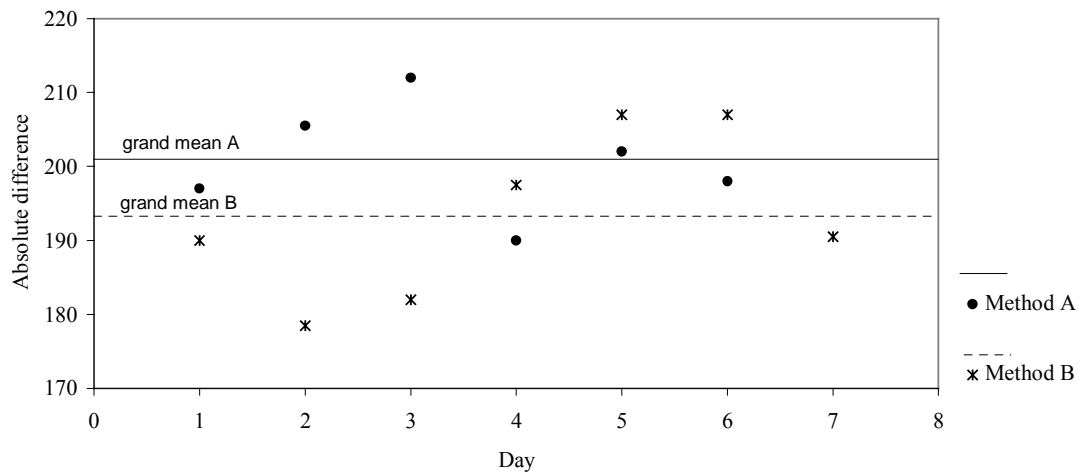
**Figure 5** Results for method A (Example 2)



**Figure 6** Results for method B (Example 2)



**Figure 7** Absolute differences between duplicates (Example 2)



**Figure 8** Day means and the grand means obtained with the two methods (Example 2)