Procedure 4

The Use of Precision Statistics

1. Scope 2

2. Definitions 2

3. CEC Precision Statistics 3

4. Measurement of a Single Product 4

5. Conformance with Specifications 6

6. Comparison of Two Products 7

7. Discriminating Power 8

8. Setting of Specifications and Safety Limits 14

8.1 Basic principles 14

8.2 Effect of test-to-test variability 15

8.3 Relative-to-reference specifications 18

8.4 Specifications involving multiple parameters or multiple tests 20

8.5 Examples of secondary “no-harm” parameters 22

9. References 24

Procedure 4

The Use of Precision Statistics

1. Scope

This procedure defines the precision statistics that are provided with CEC test methods and describes how these might be used in practice. In particular, the procedure describes how the precision of an estimate of a test parameter for a particular fluid (or the difference between two fluids), derived from one or more measurements, can be calculated from these statistics. Tables and figures are given which will allow a user to determine *a priori* the probability of measuring a significant difference between a fluid and a specified value, or a significant difference between two fluids, as a function of the unknown true difference. The procedure also describes the use of precision statistics in setting specifications and safety limits and in checking products for conformance against specifications. Absolute and relative-to-reference specifications are discussed as are specifications involving multiple parameters or tests.

The equations in this procedure are only valid for tests where within-laboratory and between-laboratory variations follow the normal distribution and do not increase or decrease as the level the parameter being measured increases. The advice of SDG should be sought for methods where the variations are non-normal or the variability is not constant.

2. Definitions

In this procedure the following definitions are used:

Actual value[1]**:** the actual quantitative value for the prepared sample (the actual value only exists for fundamental physical or chemical properties such as density, concentration, temperature, etc.)

True value: the value towards which the average of single results obtained by *n* laboratories tends, as *n* tends towards infinity

Accuracy: the closeness of agreement between a test result, or the average value obtained from a series of test results, and the true value

Bias: the difference between the true value and the actual value

Laboratory bias: the difference between the expectation of the test results from a particular laboratory and the true value

Precision: the closeness of agreement between independent test results obtained under stipulated conditions

Repeatability *r*: The value equal to or below which the absolute difference between two single test results, obtained in the normal and correct operation of the same test method on identical material, may be expected to lie with a probability of 95% when conducted under the following conditions: non-consecutive tests with intervening

[1] The actual value is referred to as the “known value” in ISO 4259 [2] which is slightly misleading as the actual value is not normally known

changes of test material, completed in a short time interval by the same operator at the same laboratory using the same apparatus[2]

Site Precision *r'*: The value equal to or below which the absolute difference between two single test results on test specimens from the same fluid batch, obtained over an extended period of time, spanning at least a 15-day interval, by one or more operators in a single site location practicing the same test method on a single measurement system may be expected to lie with a probability of 95%[3]

Reproducibility *R*: The value equal to or below which the absolute difference between two single test results obtained in the normal and correct operation of the same test method on identical material by operators in different laboratories may be expected to lie with a probability of 95%

Test result: The final value of a test parameter for a particular sample obtained by following the complete set of instructions of the test method

3. CEC Precision Statistics

The precision of a CEC test method is expressed by the *repeatability r* and *reproducibility R* for each parameter measured. The *site precision r'* is also used on occasions as a measure of long-term repeatability at a particular laboratory.

CEC uses the definitions of *r* and *R* in international standard ISO 5725 [1][4], viz.

$$r = 1.96\sqrt{2\sigma_r^2} \text{ and } R = 1.96\sqrt{2\sigma_L^2 + 2\sigma_r^2}$$

where σ_r is the *within-laboratory* or *repeatability standard deviation* and σ_L is the *between-laboratory standard deviation*.

Thus the

Within-laboratory or *repeatability standard deviation* $\sigma_r = r / 2.8$

and the

Reproducibility standard deviation $\sigma_R = R / 2.8$

Hence the

Between-laboratory standard deviation $\sigma_L = \sqrt{(R^2 - r^2)} / 2.8$.

The within-laboratory standard deviation should be estimated from the site precision *r'* when considering differences between sets of results collected some time apart.

The repeatability *r* may vary from laboratory to laboratory and so homogeneity needs to be checked when round robin data are analysed and new laboratories come on stream. In this procedure, it is assumed that a common repeatability figure *r* can be assumed for each laboratory performing the test.

---

[2] In CEC round robins, repeat tests on the same sample at the same laboratory must be conducted independently as if they were tests on different materials (see Procedure 1, section 5). Repeat tests on the same sample are not normally be conducted back-to-back, but if this is unavoidable then the full preparatory procedures required in each run of the test (e.g. flushing, recalibration, etc) must still be carried out between tests. This ensures that the repeatability estimate from the round robin provides an appropriate error estimate when comparing different fluids

[3] This procedure uses the symbol *r'* for site precision. This is more appropriate than the symbol *R'* used in ASTM D6299 [3] as site precision is best thought of as long-term repeatability at a particular laboratory. Test monitoring data are normally collected under site precision conditions.

[4] ISO 5725 [1] uses the terms "repeatability limit" and "reproducibility limit" to describe *r* and *R.*

To be accepted as fit for use for the purpose(s) defined by the Management Board, CEC methods are required to achieve appropriate repeatability and reproducibility targets and discrimination levels as described in Procedure 3.

Laboratories running CEC tests must demonstrate that they can obtain similar test result levels, repeatability and discrimination to the test development laboratory during the early stages of test introduction (see Procedure 3 section 3.5) and to the general laboratory population thereafter (see Procedure 2 section 1.14).

CEC test methods may be used to estimate the true value of a test parameter for a particular sample or to estimate the difference in true values for two or more samples. The accuracy of such estimates will depend on the experimental design, in particular the number and location of tests conducted, and the precision of the test method used.

In an ideal experiment, the precision of the test method is estimated from the variations between sets of repeat measurements on the various samples in that particular programme. However, the cost of many CEC methods, particularly engine tests, will often prohibit the collection of sufficient data to re-estimate precision. In such circumstances, the user has to rely on the precision estimates obtained in the most recent round robin.

In this procedure, it is assumed for simplicity that the repeatability *r* and reproducibility *R* are known values. In reality, these values or *r* and *R* will themselves be empirical estimates derived from round robin data and so will be subject to a degree of uncertainty. The degree of uncertainty will depend on the size of the round robin programme, which determines the number of degrees of freedom available to estimate *r* and *R*. Further details are given in Appendix B of Procedure 1. The precision statement in each CEC method shall state the number of laboratories participating in the round robin or reference testing programme, the number of samples tested and their mean values or range of mean values. Degrees of freedom and confidence limits for *r* and *R* can also be stated.

In this Procedure, it is assumed that the number of degrees of freedom is large when performing *t*-tests to determine whether parameter differences between one fluid and a specified value, or between two fluids, are statistically significant. The upper 5% point of the *t*-distribution will thus be assumed to be 1.645 for one-sided tests and 1.96 for two-sided tests. The advice of SDG should be taken if more accurate critical values of *t* (and the critical differences *CD* in Sections 4 and 6) are needed dependent on the degrees of freedom available when estimating the standard error of the difference being tested. Estimates of *R* in CEC tests are often based on fewer than 10 d.f.

4. Measurement of a Single Product

The standard deviation of a single measurement of a test parameter *X* about the true value is

$$SD(X) = \sqrt{\sigma_L^2 + \sigma_r^2} = \sqrt{\sigma_R^2} = \sigma_R$$

If a single laboratory obtains *k* independent test results on samples of a single product under repeatability conditions then the standard error of the average value $\overline{X}$ is

$$SE(\overline{X}) = \sqrt{\sigma_R^2 - \sigma_r^2\left(1 - \frac{1}{k}\right)}$$

If *N* laboratories obtain $k_1, k, \ldots k_N$ independent test results on samples of a single product then the standard error of the average $\overline{X}$ of the laboratory averages is

$$SE(\overline{X}) = \frac{1}{\sqrt{N}}\sqrt{\sigma_R^2 - \sigma_r^2\left(1 - \frac{1}{N}\left(\frac{1}{k_1} + \frac{1}{k_2} + \ldots + \frac{1}{k_N}\right)\right)}$$

The values σ_r and σ_R may be computed from the repeatability *r* and reproducibility *R* using the formulae

$$\sigma_r = r / 2.8 \quad \sigma_R = R / 2.8$$

The value $\overline{X}$ is significantly different from zero at P < 5% in a two-sided test if its absolute value exceeds the critical difference

$$CD_{5\%(2-sided)} = 1.96 \times SE(\overline{X})$$

where $SE(\overline{X})$ depends on the number of tests conducted, as detailed above.

The value $\overline{X}$ is significantly greater than zero at P < 5% in a one-sided test if $\overline{X}$ exceeds the critical difference

$$CD_{5\%(1-sided)} = 1.645 \times SE(\overline{X})$$

Similarly the value $\overline{X}$ is significantly less than zero at P < 5% in a one-sided test if $\overline{X}$ is less than $-CD_{5\%(1-sided)}$.

The critical values $CD_{5\%}$ may be computed from the repeatability *r* and reproducibility *R* using the following formulae:

(single measurement)

$$CD_{5\%(2-sided)} = \frac{R}{\sqrt{2}} = 0.71R$$

$$CD_{5\%(1-sided)} = 0.59R$$

(k measurements at 1 laboratory)

$$CD_{5\%(2-sided)} = 0.71\sqrt{R^2 - r^2\left(1 - \frac{1}{k}\right)}$$

$$CD_{5\%(1-sided)} = 0.59\sqrt{R^2 - r^2\left(1 - \frac{1}{k}\right)}$$

($k_1, k_2, \ldots k_N$ measurements at N laboratories)

$$CD_{5\%(2-sided)} = \frac{0.71}{\sqrt{N}}\sqrt{R^2 - r^2\left(1 - \frac{1}{N}\left(\frac{1}{k_1} + \frac{1}{k_2} + \ldots + \frac{1}{k_N}\right)\right)}$$

$$CD_{5\%(1-sided)} = \frac{0.59}{\sqrt{N}}\sqrt{R^2 - r^2\left(1 - \frac{1}{N}\left(\frac{1}{k_1} + \frac{1}{k_2} + \ldots + \frac{1}{k_N}\right)\right)}$$

If the value $\overline{X}$ is being compared against a specified value *A*, then

$$SE(\overline{X} - A) = SE(\overline{X})$$

where $SE(\overline{X})$ is given by the formulae above. Thus the measured difference $(\overline{X} - A)$ is statistically significant at P < 5% in a one- or two-sided test if it is greater than the appropriate critical difference $CD_{5\%}$ above.

The critical values $CD_{5\%}$ above may also be used to construct 95% confidence intervals for the true value μ of the test parameter. A two-sided 95% confidence interval for μ is

$$\overline{X} - CD_{5\%(2-sided)} \le \mu \le \overline{X} + CD_{5\%(2-sided)}$$

A one-sided 95% confidence interval for μ takes the form

$$\mu \ge \overline{X} - CD_{5\%(1-sided)} \quad \text{(lower limit)}$$

or

$$\mu \le \overline{X} + CD_{5\%(1-sided)} \quad \text{(upper limit)}$$

The value $\overline{X}$ is significantly different from a hypothesised value μ_0 at P < 5% in a one- or two-sided test if the value μ_0 lies outside the appropriate one- or two-sided 95% confidence limits constructed as above.

It can be deduced from the above equations that to improve the precision of estimation of the true value μ for a single product, it is usually more effective to take repeat measurements at different laboratories than to perform repeat tests at a single laboratory in order to reduce $SE(\overline{X})$ and $CD_{5\%}$ (unless *r* and *R* are similar).

5. Conformance with Specifications

The one-sided limits in Section 4 may be used to check whether a product conforms with specifications.

For example, a supplier who has no other source of information on the true value of a test parameter than a single test result *X* may consider that the product meets the specification, with 95% confidence, only if the result *X* is such that

$X \le A_U$ - 0.59*R* (in the case of a single upper limit A_U)

or

$X \ge A_L$ + 0.59*R* (in the case of a single lower limit A_L)

Similarly, a recipient who has no other source of information on the true value of a test parameter than a single test result *X* may consider that the product fails the specification, with 95% confidence, only if the result *X* is such that

$X \ge A_U$ + 0.59*R* (in the case of a single upper limit A_U)

or

$X \le A_L$ - 0.59*R* (in the case of a single lower limit A_L)

In each case, the margin may be reduced by performing repeat tests and replacing 0.59*R* by the appropriate value of $CD_{5\%(1-sided)}$ from Section 4. The greatest reductions are obtained when the repeat tests are conducted at different laboratories (unless *r* and *R* are similar).

6. Comparison of Two Products

If a laboratory makes single measurements X_1 and X_2 on each of two products under repeatability conditions (same operator, same apparatus, same laboratory, short intervals of time) then the standard error of the difference is

$$SE(X_1 - X_2) = \sqrt{2\sigma_r^2}$$

If that single laboratory makes k_1 independent measurements on sample 1 and k_2 independent measurements on sample 2, then the standard error of the difference in means is

$$SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\sigma_r^2 \left(\frac{1}{k_1} + \frac{1}{k_2} \right)}$$

If a set of k_1 independent measurements on sample 1 is made at one laboratory and a set of k_2 independent measurements on sample 2 is made at a different laboratory then the standard error of the difference in means is

$$SE(\overline{X}_1 - \overline{X}_2) = \sqrt{2\sigma_R^2 - 2\sigma_r^2 \left(1 - \frac{1}{2k_1} - \frac{1}{2k_2} \right)}$$

Finally if N laboratories make sets of $k_1, k_2, \ldots k_N$ independent measurements respectively on both samples then the standard error of the difference between the simple averages $\overline{X}_1$ and $\overline{X}_2$ of the $k_1 + k_2 + \ldots + k_N$ measurements on each sample is

$$SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{2\sigma_r^2}{k_1 + k_2 + \ldots + k_N}}$$

The values $\overline{X}_1$ and $\overline{X}_2$ are significantly different from each other at P < 5% in a two-sided test if the absolute value of the difference $\overline{X}_1 - \overline{X}_2$ exceeds the critical difference

$$CD_{5\%(2-sided)} = 1.96 \times SE(\overline{X}_1 - \overline{X}_2)$$

where $SE(\overline{X}_1 - \overline{X}_2)$ depends on the number of tests conducted, as detailed above.

The value $\overline{X}_1$ is significantly greater than $\overline{X}_2$ at P < 5% in a one-sided test if the difference $\overline{X}_1 - \overline{X}_2$ exceeds the critical difference

$$CD_{5\%(1-sided)} = 1.645 \times SE(\overline{X}_1 - \overline{X}_2)$$

Similarly the value $\overline{X}_1$ is significantly smaller than $\overline{X}_2$ at P < 5% in a one-sided test if the difference $\overline{X}_1 - \overline{X}_2$ is less than $-CD_{5\%(1-sided)}$.

The critical values $CD_{5\%}$ above may also be used to construct 95% confidence intervals for the true difference $\mu_1 - \mu_2$ between the values of the test parameter for the two samples. A two-sided 95% confidence interval for $\mu_1 - \mu_2$ is:

$$\overline{X}_1 - \overline{X}_2 - CD_{5\%(2-sided)} \leq \mu_1 - \mu_2 \leq \overline{X}_1 - \overline{X}_2 + CD_{5\%(2-sided)}$$

A one-sided 95% confidence interval for $\mu_1 - \mu_2$ takes the form

$$\mu_1 - \mu_2 \geq \overline{X}_1 - \overline{X}_2 - CD_{5\%(1-sided)} \quad \text{(lower limit)}$$

or

$$\mu_1 - \mu_2 \leq \overline{X}_1 - \overline{X}_2 + CD_{5\%(1-sided)} \text{ (upper limit)}$$

The critical values $CD_{5\%}$ may be computed from the repeatability r and reproducibility R using the following formulae:

(single measurement)

$$CD_{5\%(2-sided)} = r$$

$$CD_{5\%(1-sided)} = 0.84\, r$$

(samples tested at same laboratory)

$$CD_{5\%(2-sided)} = r\sqrt{\frac{1}{2k_1} + \frac{1}{2k_2}}$$

$$CD_{5\%(1-sided)} = 0.84r\sqrt{\frac{1}{2k_1} + \frac{1}{2k_2}}$$

(samples tested at 2 different laboratories)

$$CD_{5\%(2-sided)} = \sqrt{R^2 - r^2\left(1 - \frac{1}{2k_1} - \frac{1}{2k_2}\right)}$$

$$CD_{5\%(1-sided)} = 0.84\sqrt{R^2 - r^2\left(1 - \frac{1}{2k_1} - \frac{1}{2k_2}\right)}$$

(N laboratories performing k_1, k_2, … k_N tests respectively on each sample)

$$CD_{5\%(2-sided)} = \frac{r}{\sqrt{k_1 + k_2 + \ldots + k_N}}$$

$$CD_{5\%(1-sided)} = \frac{0.84r}{\sqrt{k_1 + k_2 + \ldots + k_N}}$$

The difference $\overline{X}_1 - \overline{X}_2$ is significantly different from zero at P < 5% in a one- or two-sided test if the appropriate one- or two-sided 95% confidence limits constructed as above do not contain the value 0.

The greatest reductions in the critical differences $CD_{5\%}$ above are obtained by performing repeat measurements on the two samples at the same laboratory (or the same set of laboratories). The comparison is then made within-laboratory and not between-laboratory. Thus $SE(\overline{X}_1 - \overline{X}_2)$ and $CD_{5\%}$ depend solely on the repeatability r and not the reproducibility R.

7. Discriminating Power

"Power" is the probability of measuring a significant difference between two products (or between one product and a specification) when a difference does indeed exist. Thus power is the probability of measuring values or sets of values in which the difference in means $\overline{X}_1 - \overline{X}_2$ (or the difference between the mean value for a single product $\overline{X}$ and some specified value A) is large enough to be statistically significant at P < 5%. Power is equal to

$$\text{Power} = 1 - \beta$$

where β is the probability of a type II error, that is the risk of failing to detect a difference that actually exists.

Power depends on:

(a) how large the unknown difference in true values actually is,
(b) the numbers and locations of tests conducted on each material (i.e. the experimental design),
(c) the repeatability *r* and reproducibility *R* of the test method being used,
(d) the prescribed significance level.

When the repeatability *r* and reproducibility *R* of the test method are known, the standard error of the measured difference $\overline{X}_1 - \overline{X}_2$ or $\overline{X} - A$ can be computed using the formulae given in Sections 6 and 4 respectively. The power curve for the proposed experiment can then be obtained from Figures 1 (one-sided tests) or 2 (two-sided tests), which show the probability of obtaining a significant difference as a function of the calculated standard error. These curves were computed using the formulae

$$P(\overline{X}_1 - \overline{X}_2 > 1.645) = P\left(Z < -1.645 + \frac{\Delta}{SE(\overline{X}_1 - \overline{X}_2)} \right) \quad \text{(one-sided tests)}$$

$$P(\left|\overline{X}_1 - \overline{X}_2\right| > 1.96) = P\left(Z < -1.96 + \frac{\Delta}{SE(\overline{X}_1 - \overline{X}_2)} \right) + P\left(Z > 1.96 + \frac{\Delta}{SE(\overline{X}_1 - \overline{X}_2)} \right)$$

(two-sided tests)

where *Z* is a standard normal variate with mean zero and standard deviation one and Δ is the unknown true difference (i.e. the value on the x-axis) expressed as a multiple of $SE(\overline{X}_1 - \overline{X}_2)$.

Some specifications simply require a test material to give a better test result than a reference material. Figure 3 shows the probability of obtaining a higher test result (or average test result) on product 1 than product 2 as a function of the true difference.

In Figures 1 and 2, it can be seen that the probability of measuring a significant difference at P < 5% when no difference exists is indeed 5%, the probability α of a type I error. The probability of measuring a significant difference at P < 5% is equal to 50% when the true difference is equal to the critical difference, i.e. 1.645 $\times$ $SE(\overline{X}_1 - \overline{X}_2)$ (Figure 1) and 1.96 $\times$ $SE(\overline{X}_1 - \overline{X}_2)$ (Figure 2). To have a 95% chance probability of measuring a significant difference at P < 5%, the true difference would need to exceed 3.29 $\times$ $SE(\overline{X}_1 - \overline{X}_2)$ in a one-sided test and 3.60 $\times$ $SE(\overline{X}_1 - \overline{X}_2)$ in a two-sided test.

Table 1 gives the true differences required for various stipulated power levels. For example, to have a 90% chance or better of measuring a significant difference at P < 5% in a one-sided test, the true difference would need to be at least 2.93 times the standard error of the measured difference, calculated as above.

Table 1. Power curve percentiles: the difference in true values required for the probability of the desired event to be equal to the stipulated power level. This table may be used for comparisons of two products or for comparing one product with a specification. *SE (difference)* should be computed using the appropriate formula in Sections 6 or 4.

Power level	Probability β of type II error	Desired event		
		Measured difference significant at P<5% (1-sided test) (> 1.645 × $SE(diff)$)	Measured difference significant at P<5% (2-sided test) (> 1.96 × $SE(diff)$)	Measured difference > 0
0.99	0.01	3.97 × $SE(diff)$	4.29 × $SE(diff)$	2.33 × $SE(diff)$
0.95	0.05	3.29 × $SE(diff)$	3.60 × $SE(diff)$	1.64 × $SE(diff)$
0.9	0.1	2.93 × $SE(diff)$	3.24 × $SE(diff)$	1.28 × $SE(diff)$
0.8	0.2	2.49 × $SE(diff)$	2.80 × $SE(diff)$	0.84 × $SE(diff)$
0.7	0.3	2.17 × $SE(diff)$	2.48 × $SE(diff)$	0.52 × $SE(diff)$
0.6	0.4	1.90 × $SE(diff)$	2.21 × $SE(diff)$	0.25 × $SE(diff)$
0.5	0.5	1.64 × $SE(diff)$	1.96 × $SE(diff)$	0
0.4	0.6	1.39 × $SE(diff)$	1.71 × $SE(diff)$	-0.25 × $SE(diff)$
0.3	0.7	1.12 × $SE(diff)$	1.43 × $SE(diff)$	-0.52 × $SE(diff)$
0.2	0.8	0.80 × $SE(diff)$	1.11 × $SE(diff)$	-0.84 × $SE(diff)$
0.1	0.9	0.36 × $SE(diff)$	0.65 × $SE(diff)$	-1.28 × $SE(diff)$
0.05	0.95	0	0	-1.64 × $SE(diff)$

To illustrate further the use of Figures 1 to 3 and Table 1, consider the simplest possible experiments where single measurements only are made on each sample. If the interest is in measuring the test parameter for a single sample and comparing it with a specified value, then the standard error of the difference between the single measurement *X* and the specification is

$$SE(X - specification) = \sigma_R = R/2.8$$

Table 2 gives the true differences required for various stipulated power levels as a function of the reproducibility *R*. To have a 95% chance or better of measuring a significant difference at P < 5%, in a one-sided test, that is a 95% chance of measuring a value *X* which betters the specification by at least 0.59*R*, the true difference would need to be at least 1.17 times the reproducibility *R*. If *X* is compared with a hypothesised value in a two-sided test, it will be deemed significantly different from that value at P < 5% if it differs from it by more than 0.71*R*

Table 2. Single measurement on one product - power curve percentiles: the difference between true and hypothesised values required for the probability of the desired event to be equal to the stipulated power level. These values are expressed as multiples of the reproducibility *R* of the test method.

Power level	Probability β of type II error	Desired event		
		Measured difference significant at P<5% (1-sided test) (> 0.59*R*)	Measured difference significant at P<5% (2-sided test) (> 0.71*R*)	Measured difference > 0
0.99	0.01	1.42*R*	1.55*R*	0.84*R*
0.95	0.05	1.17*R*	1.30*R*	0.59*R*
0.9	0.1	1.05*R*	1.17*R*	0.46*R*
0.8	0.2	0.89*R*	1.01*R*	0.30*R*
0.7	0.3	0.77*R*	0.90*R*	0.19*R*
0.6	0.4	0.68*R*	0.80*R*	0.09*R*
0.5	0.5	0.59*R*	0.71*R*	0
0.4	0.6	0.50*R*	0.62*R*	-0.09*R*
0.3	0.7	0.40*R*	0.52*R*	-0.19*R*
0.2	0.8	0.29*R*	0.40*R*	-0.30*R*
0.1	0.9	0.13*R*	0.24*R*	-0.46*R*
0.05	0.95	0.00*R*	0.00*R*	-0.59*R*

in either direction. For there to be a 95% or greater chance of this happening, the true value would need to differ from that hypothesised by more than 1.30*R* in either direction.

If the interest is in comparing the values of the test parameter for two samples by making single measurements on each, the best estimate of the difference between samples is obtained by taking the two samples at the same laboratory. The standard error of the difference of is then

$$SE(X_1 - X_2) = \sqrt{2\sigma_r^2} = r/1.96$$

If the tests on the two samples have to be conducted at different laboratories, the standard error of the difference is

$$SE(X_1 - X_2) = \sqrt{2\sigma_R^2} = R/1.96$$

Table 3. Single measurements of two products - power curve percentiles: the difference in true values required for the probability of the desired event to be equal to the stipulated power level. These values are expressed as multiples of the repeatability *r* of the test method. When the tests on the two samples are performed at different laboratories, *r* should be replaced by the reproducibility *R* of the test method in this table.

Power level	Probability β of type II error	Desired event		
		Measured difference significant at P<5% (1-sided test) (> 0.84*r*)	Measured difference significant at P<5% (2-sided test) (> *r*)	Measured difference > 0
0.99	0.01	2.03*r*	2.19*r*	1.19*r*
0.95	0.05	1.68*r*	1.84*r*[5]	0.84*r*
0.9	0.1	1.49*r*	1.65*r*	0.65*r*
0.8	0.2	1.27*r*	1.43*r*	0.43*r*
0.7	0.3	1.11*r*	1.27*r*	0.27*r*
0.6	0.4	0.97*r*	1.13*r*	0.13*r*
0.5	0.5	0.84*r*	1.00*r*	0
0.4	0.6	0.71*r*	0.87*r*	-0.13*r*
0.3	0.7	0.57*r*	0.73*r*	-0.27*r*
0.2	0.8	0.41*r*	0.57*r*	-0.43*r*
0.1	0.9	0.19*r*	0.33*r*	-0.65*r*
0.05	0.95	0.00*r*	0.00*r*	-0.84*r*

Table 3 gives the true differences required for various power levels as a function of the repeatability *r* or the reproducibility *R* depending on where the tests are done. To have a 95% chance or better of measuring a significant difference at P < 5%, in a one-sided test, that is a 95% chance of measuring a value X_1 which differs from X_2 by at least 0.84*r* (or 0.84*R*) in a specified direction, the true difference would need to be at least 1.68 times *r* (or *R*). If X_1 and X_2 are compared in a two-sided test, the difference will be deemed significantly different at P < 5% if it differs by more than *r* (or *R*) in either direction. For there to be a 95% or greater chance of this happening, the true values would need to differ by more than 1.84*r* (or 1.84R) in either direction.

When tests on the two samples are conducted at the same laboratory, some thought needs to be given to the time scale over which the measurements are made. It is assumed throughout this Procedure that this is similar to the time scale defining repeatability conditions in the round robin in which *r* was estimated. If the time scales are different, then the methods in Part 3 of standard ISO 5725 [1] should be

[5] CEC previously defined the discriminating power DP = 1.*84R* to be the true difference required for there to be a 95% probability of single test results on different samples at different laboratories differing by more than *R*.

used to obtain a more appropriate measure of precision. Site precision, defined is Section 2 above, is an example of such a measure.

Figure 1

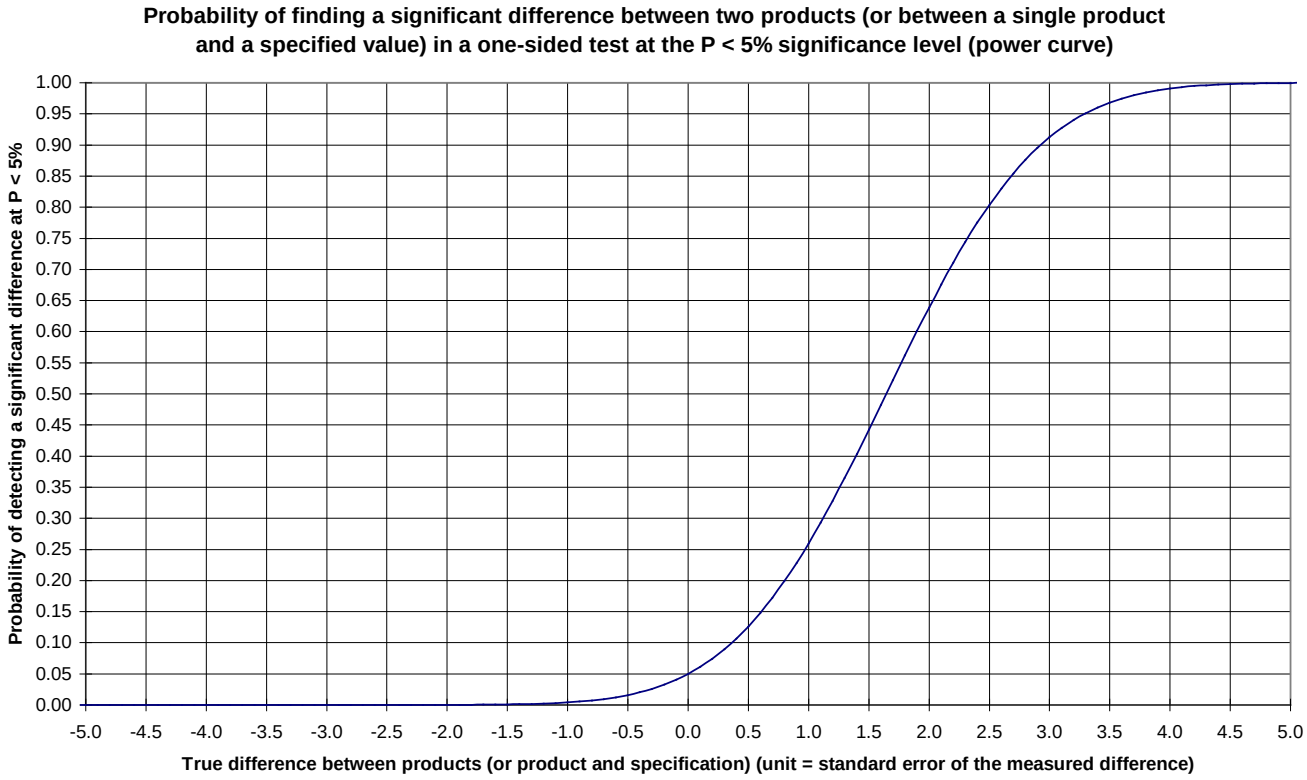


Figure 2

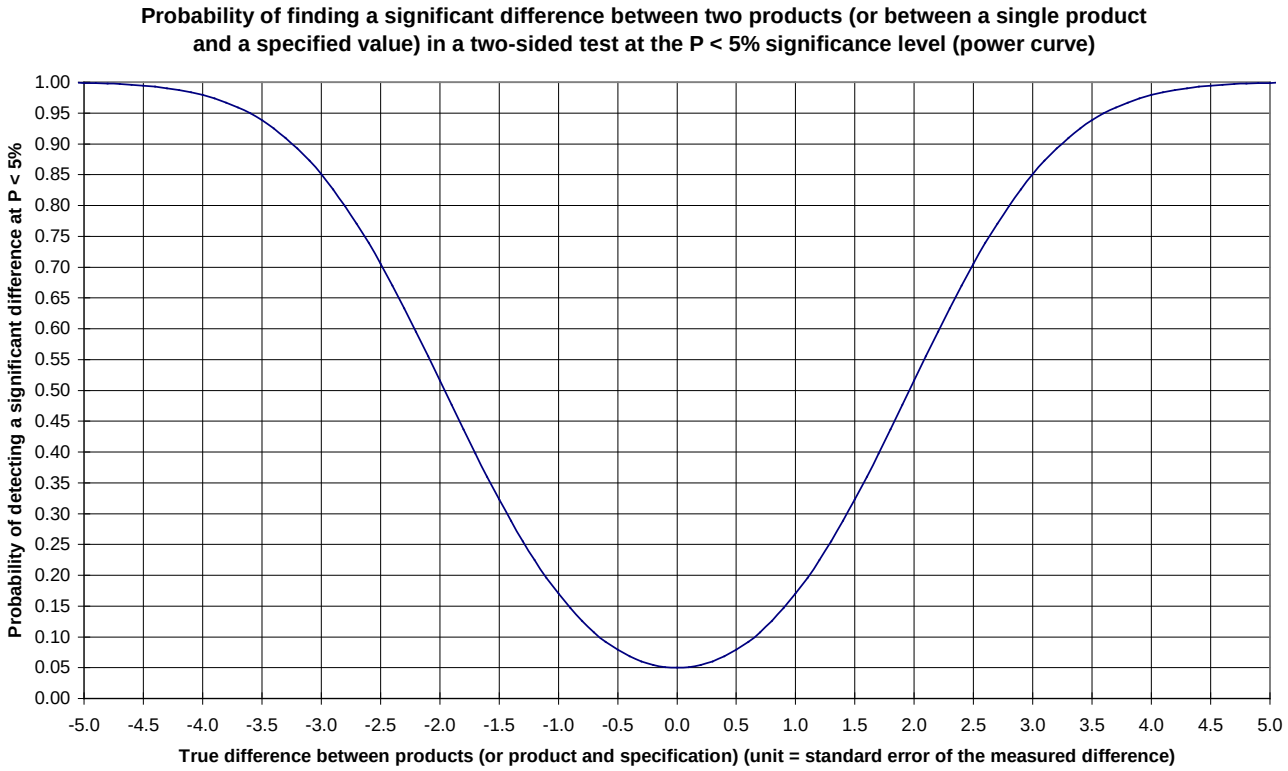
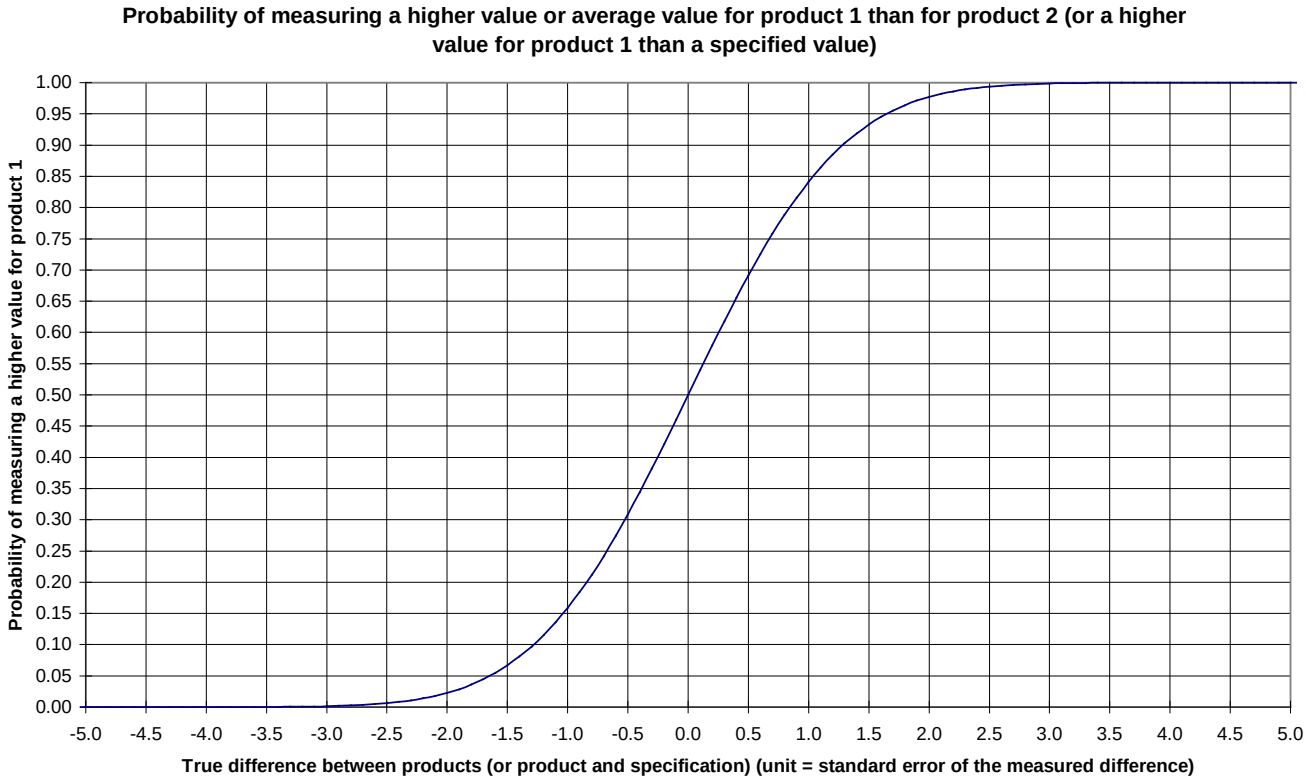


Figure 3



8. Setting of Specifications and Safety Limits

8.1 Basic principles

Specifications are limits on test parameters, set for example by regulatory authorities or motor manufacturers, which must be met before lubricants or fuels can be deemed acceptable for use in a particular context.

”Absolute” specifications can take the form of a single limit, e.g.

cam wear ≤ 10µm,

or a double limit, e.g.

tensile strength change = -40%±10%.

“Relative to reference” specifications require a candidate fluid to perform better than a reference fluid and can take a number of forms, e.g.

ring sticking test result ≤ reference result,
piston merit test result ≥ reference result – 6, or
viscosity increase test result ≤ 0.5 × reference result.

Normally the two tests will be required to take place within a short time of one another at the same laboratory to minimize the risk of drift.

The prime purpose of many specifications and safety limits is to stop fuels and oils entering the market which could

damage engines, and/or
harm the environment

Such specifications can involve secondary as well as primary parameters (as defined in Procedure 3 Section 1) and are often referred to as “safety limits”. It is not necessary for all the secondary test parameters to discriminate between reference fluids.

Other specifications are used to define different performance grades for automotive applications. For example, the 2008 ACEA sequences require Noack evaporative loss ≤ 15%wt for category A1 and ≤ 13%wt for categories A3 and A5.

Safety limits for both primary and secondary parameters are set at or below[6] the value at which the risk of harm becomes unacceptable. This requires data or scientific experience / judgement correlating CEC test results with field performance.

In many cases however, there may be little or no data available showing correlation between certain “no-harm” parameters, usually secondary measurements, and field data. The field problem of concern might be rare and most reference and candidate fluids might yield test results causing little worry. In such circumstances, there is benefit in looking at historical ranges of test results for both candidate and reference fluids. Care should be taken to examine results from a number of different manufacturers using a wide range of technologies. Some examples are shown in Section 8.5

8.2 Effect of test-to-test variability

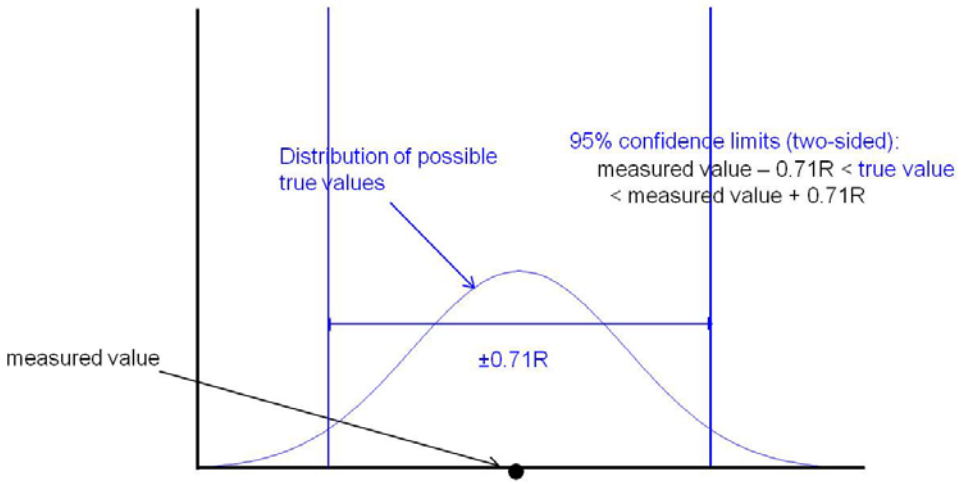
Specification setters would like to fix limits to the actual or true value (as defined in Section 2) of the parameter of interest. In practice, however, this value can rarely be established exactly and conformance checks will depend on empirical test results which are subject to test-to-test variability.

If a single measurement is made on a product, then 95% confidence limits for the true value are

Measured value +/- 0.71*R*

as shown in Figure 4.

Figure 4. Two-sided confidence interval for the true value



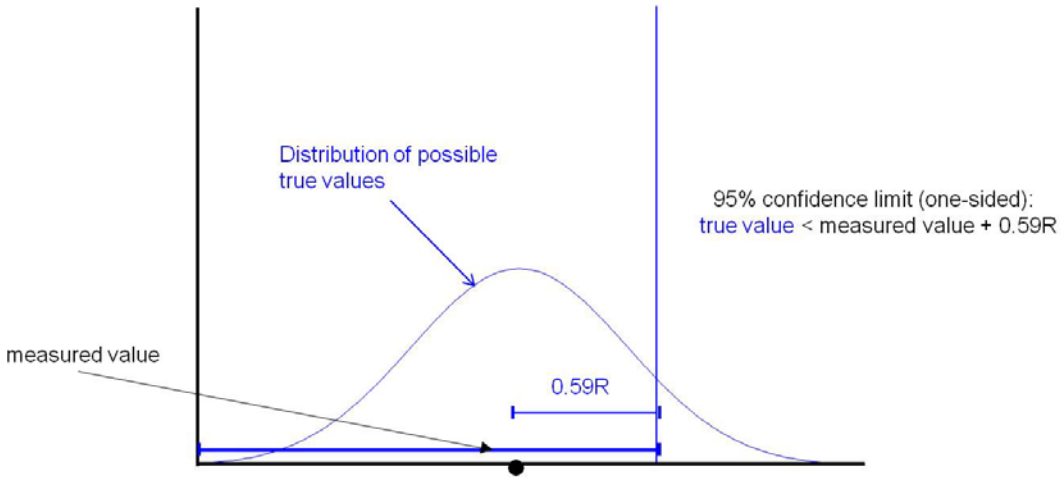
[6] Assuming harm is associated with high values

However when checking conformance with one-sided specifications, it is more useful to work with one–sided confidence limits. Thus a tester who is only worried about high values might assert with 95% confidence that the true value is less than

Measured value + 0.59*R*

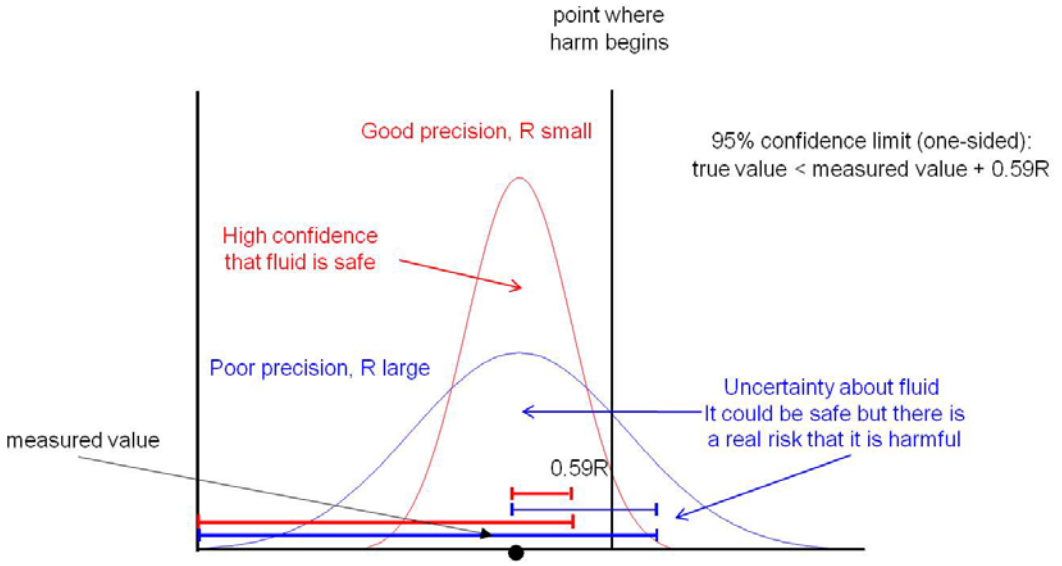
as shown in Figure 5.

Figure 5. One-sided confidence interval for the true value



It is important to have a test method with good reproducibility when testing conformance with absolute specifications. Figure 6 shows an example in which the measured value is below the point at which harm is thought to begin. When the reproducibility *R* is small then the precision is good and the tester has a high level of confidence that the product is safe. However when the reproducibility *R* is large then the precision is poor and there is a good deal of uncertainty about the true value. While the product could indeed be safe, there is also a real risk that it could be harmful.

Figure 6. Effect of reproducibility on our confidence that a fluid is safe



International Standard ISO 4259 [2] suggests a starting point as regards how good the reproducibility R needs to be. It states that for a 2-sided specification of the form $A_L \leq X \leq A_U$, the specified range $A_U - A_L$ shall not be less than $4R$ where R is the reproducibility of the test method adopted. Thus if the range is fixed then R needs to be $\leq (A_U - A_L)/4$.

If the specification can only be failed in one-direction (e.g. if there is an implied lower bound of zero in a specification of the form $X \leq A_U$), then the specified range (stated or implied) shall not be less than $2R$. Thus if the limit A_U is fixed then R needs to be less than $A_U/2$.

The reproducibility R determines the specification limit or range that the test method is capable of supporting. However it would clearly be imprudent to set a specification limit based on $2R$ or $4R$ if R is so large that this would risk harmful product reaching the field. In such cases, the test method (or parameter) is not fit for purpose and should not be approved by the CEC Management Board. The aim should be to improve the precision of the test method by setting a reproducibility target based on the desired limit, e.g. the point where harm begins, and, for example, the $2R$ or $4R$ rule (see Procedure 3, Appendix C). In some cases, a combination of improved precision and widened limits might be appropriate.

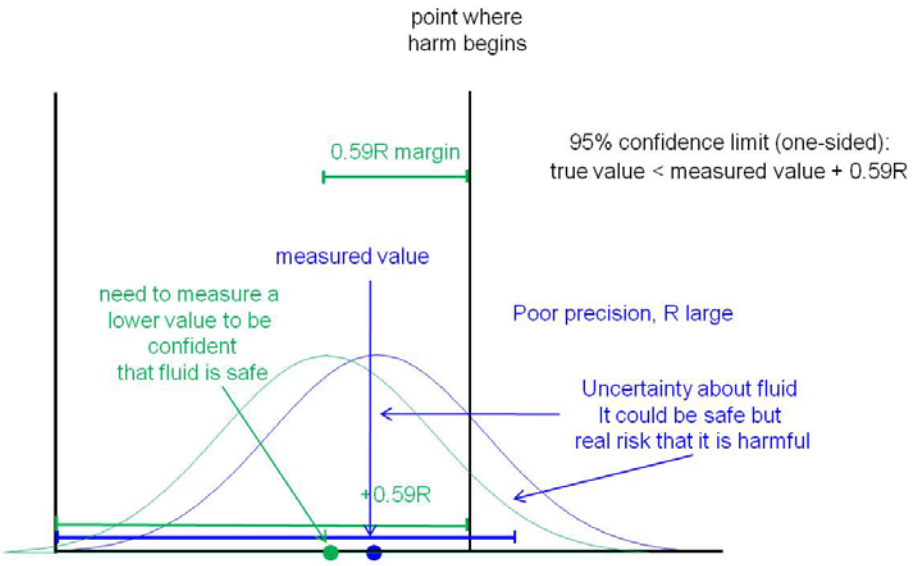
Poor test method reproducibility also increases the risk of acceptable products failing the specification due to test-to-test variability, leading to unwarranted retesting and/or re-engineering costs for the producer. A balance therefore needs to be drawn between the risk of good products failing and bad products passing the specification. Issues which need to be taken into account include (a) the precision of the test method or parameter, (b) how critical the parameter is, (c) how well the point of first harm is known, (d) the candidate retesting rules and (e) the number of parameters mentioned in the specification.

Specifications are not always set at the point at which a required level of performance is achieved or, conversely, harm is thought to begin. In the example in Figure 6, when the precision is poor (R large), the tester would need to measure a lower value for all parties to be confident that the fluid is safe (see Figure 7). If the parameter is critical, the specification setter might be inclined to introduce a “margin” to cater for test-to-test variability and thus set the specification limit lower than the desired threshold. A margin of $0.59R$ is required for 95% confidence that the true value is below a particular bound.

Margins might also be introduced to account for uncertainty in our knowledge of the point of first harm. However lowering the specification limit in this way increases the risk of good products, which cause no harm, failing the specification. Therefore a careful balance needs to be made between the two kinds of risk.

Margins might also be introduced to compensate for rules which allow the retesting of candidates since these can increase the risk of a bad candidate passing the specification.

Figure 7. Margin needed for 95% confidence that a product causes no harm



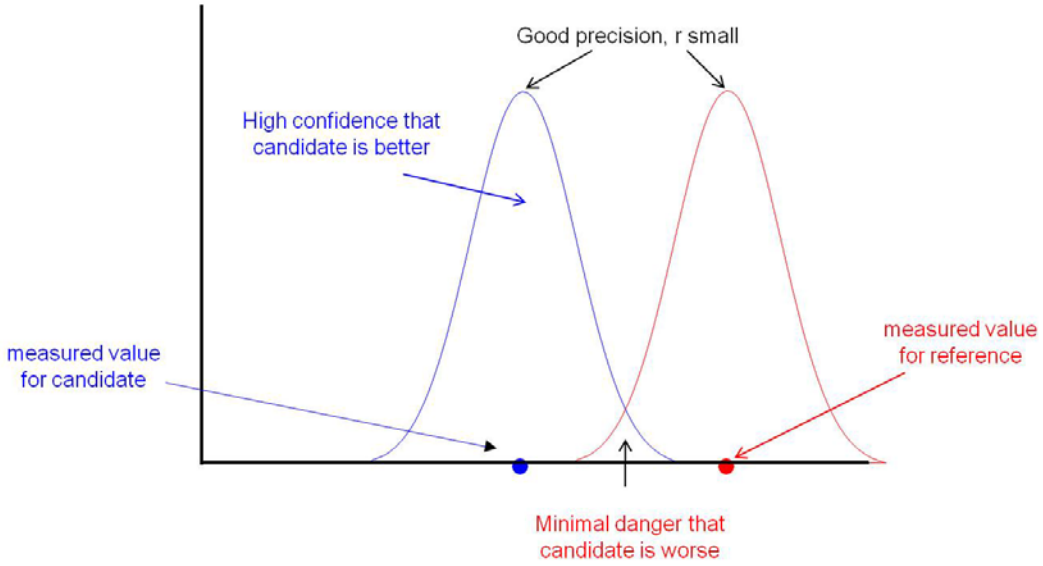
8.3 Relative-to-reference specifications

"Relative to reference" specifications do not set absolute limits A_L or A_U but require a candidate fluid to perform better than a reference fluid and can take a number of forms, e.g.

ring sticking test result ≤ reference result,
piston merit test result ≥ reference result – 6, or
viscosity increase test result ≤ 0.5 × reference result.

Tests need to have good repeatability if they are to be used for checking conformance with relative specifications. Figure 8 shows that when the repeatability is small then there is little doubt as to which is the better fluid.

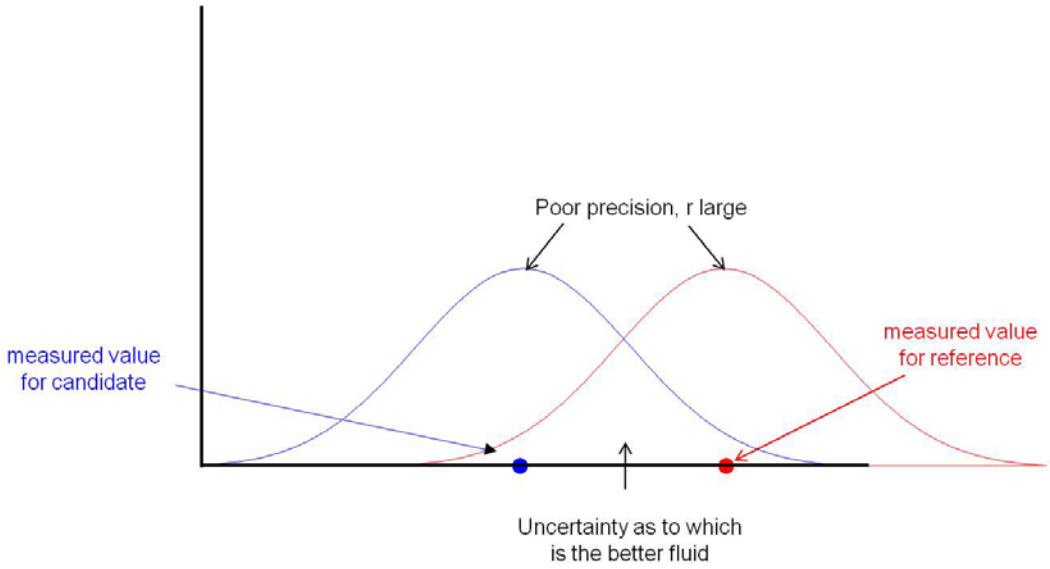
Figure 8. Discrimination between candidate and reference fluids when the repeatability *r* is small.



If the repeatability is poor, then there will be a good deal of uncertainty when the candidate and reference give similar results. This is illustrated in Figure 9.

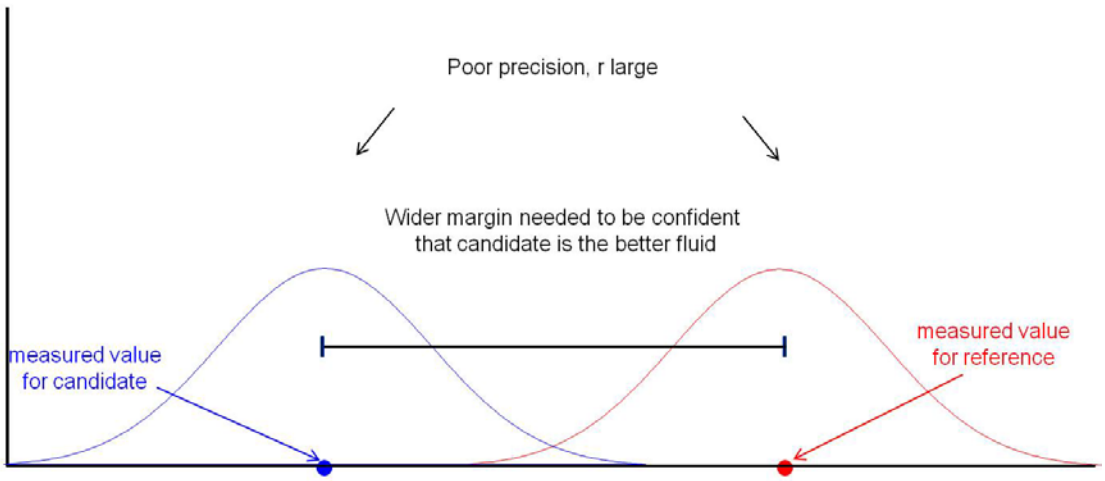
Appendix A of Procedure 3 describes how appropriate repeatability targets might be set for a parameter which is going to be used in a relative-to-reference specification.

Figure 9. Discrimination between candidate and reference fluids when the repeatability *r* is large.



In the example in Figure 9, when the precision is poor (*r* large), the tester would need to measure a larger difference between the two fluids for all parties to be confident that the fluid is safe (see Figure 10).

Figure 10. Margin required for 95% confidence that the candidate fluid is better than the reference.



If the parameter is critical in the sense that the candidate must be no worse than the reference, the specification setter might be inclined to introduce a “margin” to cater for test-to-test variability and require the candidate to beat the reference by this margin in order to reduce the risk of a poor product passing the specification. Thus the specification might take the form

Candidate result < Reference result – x

or

Candidate result < k × Reference result (where $k < 1$)

A margin of 0.84r is needed for 95% confidence that the candidate is the better fluid.

Specifications of the above form might also be used to define different performance levels.

Margins might also be introduced to account for uncertainty in our knowledge of the point of first harm and how good the reference is perceived to be. However lowering the specification limit in this way increases the risk of good products, which cause no harm, failing the specification. Therefore a careful balance needs to be made between the two kinds of risk.

Ideally the candidate and reference fluids ought to be tested consecutively within a short time of one another. Indeed some tests do involve the direct comparison of the two fluids. However some tests are so expensive that it is not realistic to perform a reference test immediately before every single candidate test. In such circumstances, several candidates may be compared against the most recent reference (or the average or moving average of recent references). The method will then require good “site precision”, that is longer-term repeatability (see Section 2 for definition). When determining the maximum number of candidate tests and elapsed time to be allowed before a new reference has to be conducted, the specification setter should take into account historical test monitoring and site precision data.

Absolute specifications are normally preferred to relative-to-reference specifications as only one test is required and the result is only subject to experimental error from that one test. However in tests where there is a high level of laboratory-to-laboratory variability, e.g. when the reproducibility R exceeds the repeatability r by a factor rather more than 1.4 (i.e. $\sqrt{2}$)[7] or more, then relative-to-reference specifications might be preferred. Examples are tests where laboratories might have particularly mild or particularly severe installations. The laboratory bias cancels out when two results are compared. Relative-to-reference specifications might also be considered in exceptional situations where the test performance threshold is uncertain in absolute terms, but the reference fuel perceived to be borderline. The threshold will often be uncertain in the early stages of test development process when the method is only installed in a limited number of laboratories.

8.4 Specifications involving multiple parameters or multiple tests

The risk of acceptable products failing specifications is particularly acute when these involve multiple parameters, or indeed multiple tests as for example in the ACEA sequences. The exact risk is difficult to quantify as this will depend on (a) the true value of the various test parameters, (b) the precision of the various test methods and (c) the correlation between the various measurements. The risks are

[7] The standard error of the difference between two measurements is lower than the standard deviation of a single measurement if $R/r > \sqrt{2}$. However any precision gain must be offset against the cost of the extra measurement.

exacerbated when specifications include parameters with poor precision. The precision of all parameters used in a specification, be they primary or secondary, should be evaluated via round robins and/or test monitoring.

Figure 11. Specification based on two correlated factors

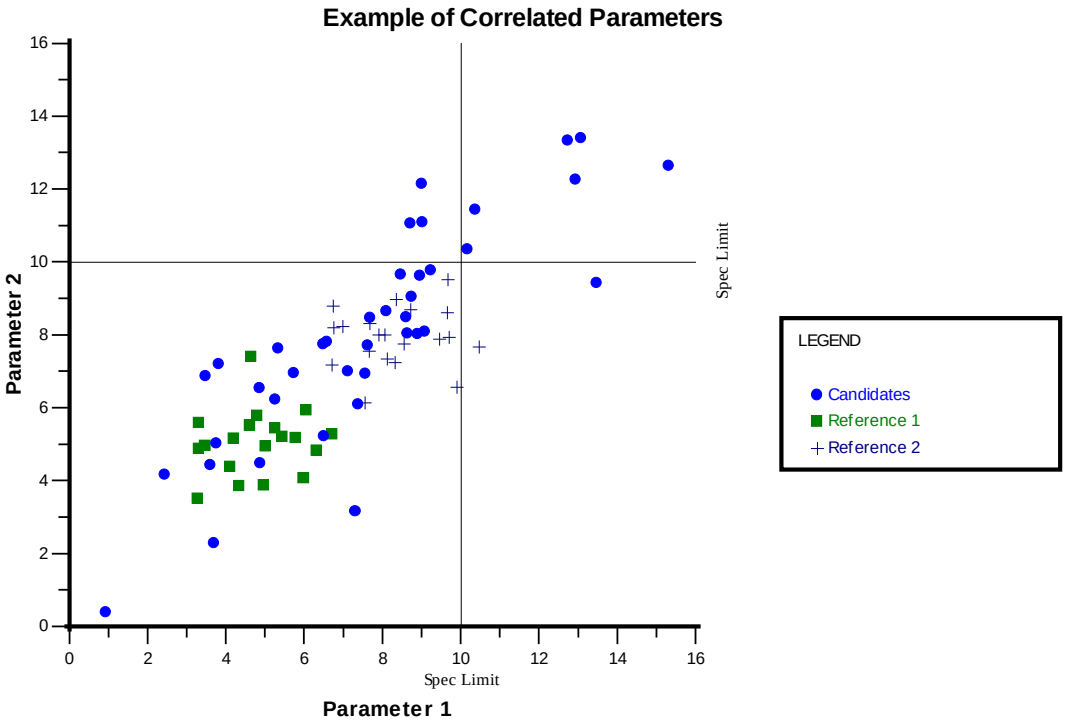


Figure 11 shows an artificial example with two correlated parameters and a specification limit of 10 units maximum on each. Seven candidates fail the specification on Parameter 1, if this is examined first, then another three subsequently fail on Parameter 2. Similarly if Parameter 2 is examined first then nine tests fail due to Parameter 2 and one further test fails due to Parameter 1. Thus the inclusion of both parameters reduces the probability of passing the test. It could of course be that all 10 failing candidates could potentially cause problems and so deserve to fail. On the other hand, some of the 10 may be safe and the failures may be due to measurement error and the increased risk engendered by the second parameter. Therefore there is an argument for using just one of these correlated parameters in the specification – or basing the specification on their sum or average if these are measuring similar phenomena in the same units.

While Figure 11 is an artificial example, there are a number of real CEC tests with highly correlated parameters. For example, in the OM646LA cam wear test, both outlet and inlet wear are measured. The two measures are highly correlated but outlet wear has better precision and discrimination. Therefore the WG and Management Board decided only to approve outlet wear as a primary parameter. Using the average or sum of the two cam wear measures was considered as an alternative but was discarded because of the high variability in inlet wear.

It is worth noting that the inter-parameter correlation in Figure 11 is less apparent when we just look at the reference fluids as these cover a narrower performance range. Therefore the degree of correlation might be underestimated in round robin or test monitoring exercises.

8.5 Examples of secondary “no-harm” parameters

This section considers a number of examples of measurements which might be considered as secondary “no-harm” parameters and discusses which of these might be suitable for approval by the CEC Management Board. The parameter needs to be considered in tandem with the likely safety limit.

Figure 12 shows the cylinder liner wear values measured for the two reference oils and a large number of candidates in the OM646 diesel engine wear test. The candidate data in Figures 12 and 13 are taken from the ATC-ERC data base and are shown by their kind permission.

Cylinder liner wear could be approved is a secondary/safety parameter in this test. The upper safety limit of 5.0 microns maximum set by ACEA is higher than all the test and reference results and so is unlikely at present to pose a serious risk to producers or consumers. And the reproducibility *R* is 1.9, which is lower than 5.0/2 = 2.5, and so conforms to ISO 4259 guidelines. This specification provides protection to the vehicle manufacturer in the event that a new oil is developed which causes significantly higher levels of cylinder liner wear.

Figure 12. Cylinder liner wear in the OM646 diesel engine wear test

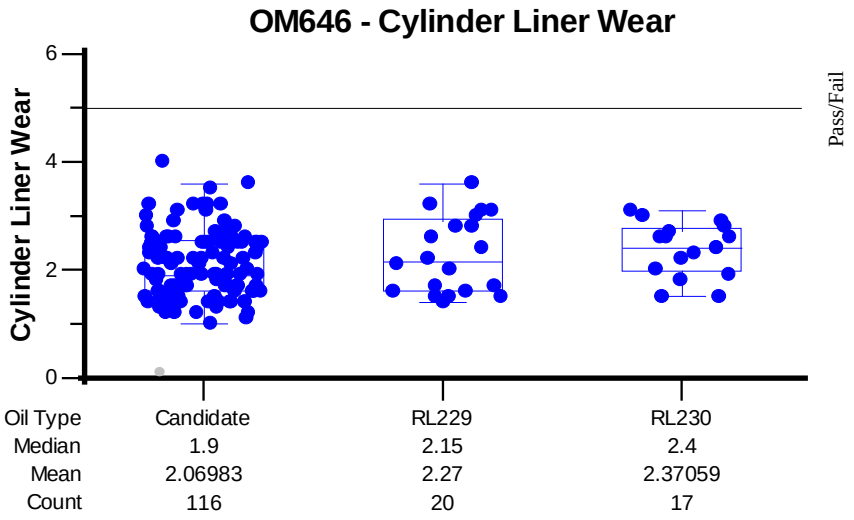


Figure 13 shows the maximum bore polish measurements from the same test. The reference results are highly scattered relative to the specification width (and interestingly worse than most of the candidates), indicating that the reproducibility is very poor. A specification limit of 3% cannot be entertained without a huge improvement in precision. If the specification really needs to be set at 3% to avoid the risk of harm, then the maximum bore polish parameter is unfit for purpose and should not be approved. As the reproducibility *R* for the reference oils is about 5.4, the 2*R* rule suggests that the method would only support a limit of 11% or higher.

Figure 13. Maximum bore polish in the OM646 diesel engine wear test

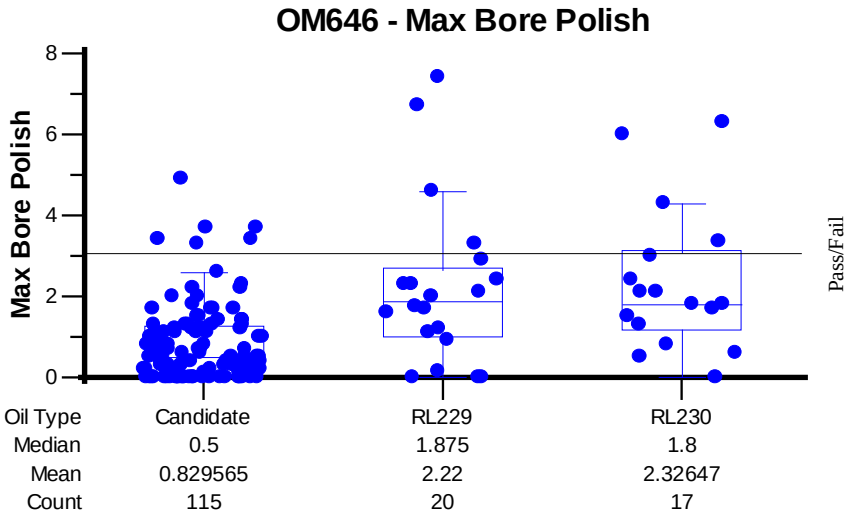


Figure 14. Test results on a potential secondary “no-harm” parameter

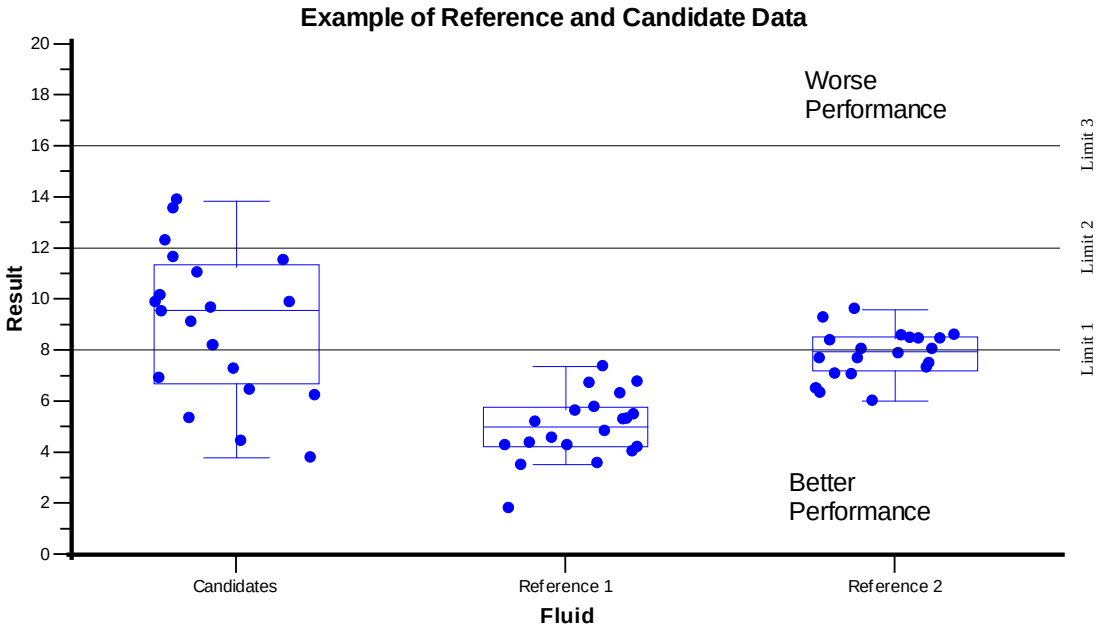


Figure 14 shows a hypothetical example of data obtained on a secondary “no-harm” parameter, which takes non-negative values. In this case there are some performance differences between the fluids; however there is poor discrimination between the reference fluids.

The fitness for purpose of this measurement as a “no-harm” parameter will depend on where the safety limit is likely to be set. In this example R = 2.8, so ISO 4259 suggests that the safety limit should be at least 2 × 2.8 = 5.6. Suppose that through consideration of the value at which the risk of harm becomes significant, the limit is required to be set at 16 units. This comfortably meets the ISO 4259 requirement and all the candidates and references pass the test. Therefore test parameter is fit for purpose as a secondary parameter, similar to cylinder liner wear in Figure 12.

If, on the other hand, the point of first harm is believed to be 8 units and the limit is set there, then the specification still meets the 2R rule but many of the Reference 2

results would be failures. This raises a number of concerns. While reference fluids are often chosen to be of borderline performance vis-à-vis primary parameters, they are normally expected to produce passes on “no-harm” parameters. If reference 2 is a realistic fluid, then significant numbers of candidates might fail an 8-unit limit in addition to the references, as indeed is the case in this example. This particular parameter thus becomes of critical importance in passing the test. It is clearly undesirable to have a specification involving a secondary/safety parameter that produces such a high number of failures and this should not be approved.

If an intermediate value (e.g. 12 units) is proposed as the safety limit, then there will need to be a more detailed consideration of the situation. In the example in Figure 14, a small number of candidates fail this limit but all the reference results are passes. The parameter might still be judged suitable as a safety parameter if it is reasonable to expect a small number of candidates to be potentially harmful and thus fail. If this is not the case, then there may need to be an improvement in test precision and/or an adjustment to the limit.

9. References

[1] International Standard ISO 5725. Accuracy (trueness and precision) of measurement methods and results.

[2] International Standard ISO 4259. Petroleum Products - Determination and application of precision data in relation to methods of test.

[3] ASTM Standard ASTM D6299. Standard Practice for Applying Statistical Quality Assurance Techniques to Evaluate Analytical Measurement System performance.