Procedure 3

Statistical Requirements for CEC Test Methods

Procedure 3

Statistical Requirements for CEC Test Methods

1. Purpose

The Management Board shall define the function and purpose of each CEC test. The function might be to measure the tendency of a particular fluid to cause (or alleviate) a particular problem. The prime purpose might be to estimate a fuel or lubricant performance measure to a given degree of accuracy, to discriminate between fuels or lubricants showing different levels of field performance, or to set fuel or lubricant specifications and subsequently check conformance.

CEC test methods may report several different parameters. Multiple “primary” parameters might measure fuel or lubricant performance in a different way (e.g. piston deposits and piston rating). The MB might also approve “secondary” parameters serving either a similar or a different purpose. The former might include less critical measures, such as piston ratings in an engine cleanliness test, which are important in their own right, but do not feature in specifications. The latter might include “no-harm” measures such as bore polish in a valve train wear test, cylinder liner wear in an engine cleanliness test or conversely piston deposits in a wear test.

To be fit for use for the purpose(s) defined by the Management Board, CEC test methods must achieve appropriate statistical precision and discrimination levels for the primary parameters. This procedure defines the process for setting guidelines and targets for precision and discrimination that must be achieved at each stage of the test development process for a test method to gain acceptance from the CEC Management Board. Precision and/or discrimination guidelines and targets might also be set for secondary parameters to ensure that they are fit for their intended purpose; secondary “no-harm” parameters might not be required to discriminate between reference fluids if these are not expected to cause problems and meet prescribed safety limits. The precision is summarized in a Precision Statement in section 11 of the test method.

2. Overview

The test development process is outlined in Guideline 9. The eventual aim is to establish methods which can be used by any accredited laboratory to obtain accurate estimates of fuel or lubricant performance, or differences in performance, and to check conformance with specification limits. Test methods thus require good precision (repeatability r and reproducibility R; see section 2 of procedure 4 for definitions) and discrimination.

The normal statistical requirement is for the working group to agree a reproducibility target for each primary parameter which must subsequently be achieved in a round robin with a minimum of 5 laboratories. However when a test is to be used primarily to compare fluids and/or to check conformance with 'relative to reference' specifications, the test will need to achieve an agreed repeatability target.

Working groups might also set precision targets for secondary parameters to ensure that they are fit for their intended purpose. Safety limits might also be set for “no

harm" parameters which the various reference fluids would be expected to meet on a consistent basis. However there may be no requirement to discriminate between reference fluids if these are all expected to pass a specification or safety limit.

The design of round robin studies and the calculation of repeatability and reproducibility are described in Procedure 1.

Individual test laboratories will need to demonstrate that they can achieve repeatable results similar to their peers. Guidelines for assessing the performance of laboratories during the early stages of test introduction may be found in Appendix B.

For each test method development, the initial statistical guidelines or targets for repeatability and discrimination will be defined in the tender document. The repeatability and discrimination statistics achieved by the test development laboratory in phase 1 shall be reported to the Management Board. The Management Board will then decide whether to accept the test for roll-out to further laboratories (phase 2), refer the test back to the Test Development Group (TDG) for further development or reject the test. At the end of phase 1, the Management Board may also declare the test fit for use, by the test development laboratory alone, for comparing fluids. This could include the checking of conformance with relative specifications in which the performance of a candidate fluid is compared with that of a reference fluid.

In phase 2 of the test development phase (see Guideline 9), the method is published and then rolled out to all laboratories participating in the TDG for the first round robin. The TDG will propose appropriate reproducibility targets for this round robin having taken advice from the SDG LO. These will then be forwarded to the Management Board for endorsement.

Initially a small pilot programme may be conducted, involving a limited number of laboratories, to check the portability of the test and the levels of laboratory-to-laboratory variability. A full round robin will then ensue. See Procedure 1 for details.

At the completion of the round robin, the TDG will compare the reproducibility (and/or repeatability) actually achieved against the respective targets and report back to the Management Board. The Board will then decide whether or not to accept the test.

On some occasions, a test will meet its targets for some parameters but not for others. In such circumstances, the Board may grant partial acceptance of the test, approving certain parameters but not others; at least one primary parameter must be approved. In such circumstances, the precision statement should clearly set out the capability of the test vis-à-vis each reported parameter. Procedure 4 describes how the capability of the test can be determined from its reproducibility and repeatability. The test method should clearly state which parameters are approved. Parameters which do not meet their targets as primary parameters could nevertheless be approved as secondary parameters if they meet the necessary requirements.

Once the test method has been accepted, the TDG and Phase 2 will be signed off and a Surveillance Group (SG) will be formed. The SG will continue to monitor its precision to maintain the quality of the test. This will normally be achieved using a test monitoring scheme which shall be developed by the SG in collaboration with SDG and approved by the Management Board (see Procedure 2). New laboratories will be required to demonstrate that they can achieve acceptable results and

repeatability using the procedures identified in Procedure 2 section 1.14. However further round robins may need to be conducted if the Board or SG consider that the test or its precision or the fluids it is used upon are likely to have changed. The Surveillance Group will also be expected to try and improve the precision and/or discrimination of those test parameters that have failed to achieve their targets and have not yet been approved.

3. Processes

i) Parameters

a) The Technical Development Group and Management Board shall agree upon the parameter(s) to be reported in the test at an early stage of the test development process. “Primary” parameters reflect the prime purpose of the test and are usually measures of fuel or lubricant performance (e.g. piston deposits, piston rating, cam wear, Noack evaporative loss, …).

Methods can also have “secondary” parameters serving either a similar or a different purpose. The former might include secondary measures such as piston ratings (as opposed to deposits) in an engine cleanliness test, or vice versa, or percent viscosity loss (as opposed to absolute final viscosity) in a shear stability test, which are important in their own right but are less critical to fluid acceptance and do not feature in ACEA specifications. The latter might include “no-harm” measures such as bore polish in a valve train wear test, cylinder liner wear in an engine cleanliness test or conversely piston deposits in a wear test. “No-harm” parameters can appear in specifications as “safety limits”.

Reference fluids are normally expected to meet safety limits on a consistent basis and there may be no requirement to discriminate between them if this is the case. Most candidate fluids would also be expected to meet safety limits and where possible this should be verified using relevant databases. Nevertheless “no-harm” parameters should be capable of identifying fluids which might cause problems in this respect, e.g. a new formulation technology might cause problems where none had been experienced previously. Note: parameters such as oil consumption would normally be considered as indicators of operational validity and would not be considered as primary or secondary parameters.

b) In tests where multiple parameters are reported, missing parameters will not necessarily invalidate other parameters in that test. However operational parameters outside tolerance limits or operational faults will normally invalidate the test and all its parameters.

c) Careful thought needs to be given to the impact of multiple parameters on laboratory acceptance (see Appendix B), test monitoring (see Procedure 2) and the setting of specifications (see Procedure 4 Section 8). It is generally undesirable to have multiple parameters unless there are strong reasons to do so with each serving a specific and different purpose as these introduce complexity in all of these areas. Thus highly correlated and physically related parameters should not be approved (but might be averaged or summed if on

the same scale). Interestingly Appendix K of the ACC code of practice requires that there be no more than a 0.85 correlation between reported parameters.

ii) Repeatability/Discrimination

a) During the initial establishment of the tender – or the setting up of a multi-laboratory development, the Management Board shall, in consultation with such technical experts as required, involve the SDG in designing a suitable series of experiments (phase 1) to demonstrate the repeatability and discrimination of the new test. The tender should allow the flexibility to run pilot tests for methods that are not fully defined at the tender stage and need refinement by the test development laboratory.

b) The Management Board, technical group and SDG shall develop guidelines for the repeatability and discrimination required of the new test to be published in the tender document. If little or no test data is available on the new method, the guidelines may be based on the ability to discriminate between particular fluids. Guidelines for setting the repeatability and discrimination targets are given in Appendix A.

c) During the test development process, the SDG liaison officer shall work with the TDG Chairman to ensure these guidelines are met and suitable analysis included in the final report. This will include an assessment of the capabilities of the test.

d) On the basis of this report, the Management Board will then decide whether to accept the test for roll-out to further laboratories (phase 2), refer the test back to the TDG for further development or reject the test. At the end of phase 1, the Management Board may also declare the test fit for use, by the test development laboratory alone, for comparing fluids (see Appendix A).

iii) Pilot inter-laboratory programme(s)

a) When phase 1 of the test development stage is complete, and the Management Board has approved progression to phase 2 (multiple laboratory development), a pilot programme may be conducted at a small number of new laboratories to

- verify the operational details of the test and that operators can follow the test procedure
- check sample distribution and handling procedures
- roughly estimate laboratory-to-laboratory variability and repeatability at other laboratories

b) Before embarking on a full round robin, the Test Development Group may also elect to conduct a mini round robin in which laboratories perform one test on one or two samples in order to obtain a preliminary estimate of reproducibility across a wider population of laboratories.

iv) Reproducibility

a) Once the phase 1 (single laboratory) test development is approved by the Management Board, and any pilot inter-laboratory programmes are complete, a round robin shall be conducted at a wider population of laboratories to determine the reproducibility of the test method and to confirm its repeatability

at other locations (phase 2). See Procedure 1 for details on the design, conduct and analysis of round robins.

b) To be accepted as fit for the purpose defined by the Management Board, a CEC test parameter must achieve a repeatability $r \leq r_{target}$ and/or a reproducibility $R \leq R_{target}$ in a round robin, where r_{target} and R_{target} are set by the Management Board, TDG and the SDG liaison officer as described in Section 4.2. Guidelines for setting the repeatability and reproducibility targets are given in Appendix C.

c) In situations where there are less than 5 laboratories with equipment installed, it will not be possible to estimate the reproducibility of the test method to an acceptable degree of precision. In such circumstances, the approach will be to compare test result levels, precision and discrimination at each new laboratory against those at the test development laboratory (see Appendix B).

d) The responsibilities for setting repeatability and reproducibility targets are as follows:

a) Management Board
- Approve the repeatability and reproducibility targets based on the recommendations of the TDG and SDG Liaison Officer.

b) TDG Chairman
- Initiate the repeatability and reproducibility target setting process at the appropriate time (see section 4.1). Targets will normally be agreed at TDG meetings.
- Organise round robins to generate the data.

c) SDG Liaison Officer
- Provide advice on setting repeatability and/or reproducibility targets based on the purpose(s) of the test, as defined by the Management Board, and the requirements of TDG members.
- Statistically analyse the round robin data to determine
  $Q_r = r / r_{target}$ and/or $Q_R = R / R_{target}$
  for each reported parameter and thus advise whether the precision achieved meets requirements.

c) Participating Laboratories
- Participate in round robins as required.

v) Initial acceptance of laboratories

a) The first tests outside the test development laboratory (phase 2) will serve a dual purpose. They will help assess both the portability of the test and also the ability of the participating laboratories to obtain similar results to the test development laboratory.

b) During the early stages of test introduction, new laboratories may be considered to meet CEC's quality requirements, as given in Guideline 18, if they achieve similar test result levels and similar repeatability and discrimination to the test development laboratory. Guidelines may be found in Appendix B.

c) Laboratories participating in the first round robin may be considered to meet CEC's quality requirements (Guideline 18) if they achieve similar test result levels and similar repeatability and discrimination to the other participants as determined by the guidelines given in Appendix B. New laboratories introducing the test after the first round robin will be compared against the general population using the procedure in Procedure 2 section 1.14.

4. Process for Determining Repeatability and Reproducibility Targets

4.1 Timing

At the start of the test development phase, the Management Board, Technical Development Group and SDG shall develop guidelines for the repeatability and discrimination required of the new test to be published in the tender document as per Section 3(i) above.

The process for determining the repeatability and reproducibility targets for a new method in its first round robin should be initiated by the TDG Chairman with the approval of the Management Board at the start of phase 2. Ideally the test method should be in a stable state with no further major changes envisaged which would impact the setting of reproducibility targets. The repeatability and reproducibility targets should only be finalised once the pilot programmes (if any) are complete.

The repeatability and reproducibility targets will need to be reviewed if changes are subsequently made to the test method that could have a major impact on the test results (e.g. increased or decreased severity) (see section 4.3).

4.2 Process and participants

The determination of repeatability and reproducibility targets is the responsibility of the TDG, as they have the experience of the test. The TDG should work with its Statistical Liaison Officer to propose a recommendation to the Management Board. This proposal will then be considered by the Board.

The repeatability and/or reproducibility targets will normally be agreed at working group meetings. Guidelines for Setting the Repeatability and Reproducibility Targets may be found in Appendix C.

4.3 Updating the Repeatability and Reproducibility Target(s)

The usage of the test method and the repeatability and reproducibility targets shall be reviewed from time to time by the TDG/SG and the Management Board.

In particular, a review will be required if changes are made to the test method which have a substantial impact on the test results (e.g. increased or decreased severity), or if the test method is subsequently used in ways that were not originally envisaged. The TDG/SG Chairman should inform the Management Board in such circumstances and seek their approval for any changes to the repeatability and/or reproducibility targets.

5. Demonstrating the Reproducibility Target(s)

In order to demonstrate compliance with the requirements, a full round robin of all laboratories participating in the working group should be carried out, as described in Procedure 1. A minimum of 5 laboratories is required for the full round robin. If

fewer than 5 laboratories are available the procedure described in Section 3(iv)(c) shall apply.

The precision (repeatability and reproducibility) shall then be determined using the methods described in Procedure 1 section 8.

6. Comparison of Test Methods

In the early stages of development of a new test, there may be a number of competing test methods or variants of a test method. It is not appropriate to compare these using

$Q_r = r / r_{target}$ or $Q_R = R / R_{target}$

as repeatability and reproducibility targets would not normally be available. Instead, the methods may be compared by performing pilot precision studies on the same set of reference fluids. The following statistics may be calculated to compare repeatability, reproducibility and discrimination.

Signal to Noise for Repeatability (Higher is Better)

$$S/N_r = \frac{\text{SD of Sample Means}}{\text{Repeatability SD}}$$

Signal to Noise for Reproducibility (Higher is Better)

$$S/N_R = \frac{\text{SD of Sample Means}}{\text{Reproducibility SD}}$$

Discrimination (Lower is Better)

For each pair of samples where the performance of one fluid is known to be superior to that of the other calculate:

$$\frac{R}{\Delta} = \frac{\text{Reproducibility}}{\text{Difference in means}}$$

(alternatively use $DP/\Delta = 1.84R/\Delta$)

See Procedure 4 for definitions.

7. Precision Statement

When a CEC designated test method is published, its precision shall be detailed in a precision statement in Section 11. This will normally be based on the results of most recent Round Robin(s), and be approved by the SDG liaison officer responsible for the test. Test monitoring data can also be used to estimate Reproducibility.

Procedure 4 describes how the statistical performance of the test method in measuring and comparing fluids, and in setting and checking conformance with specifications, can be determined from the precision statement. This must meet the requirements below.

The precision statement shall state

- The source of the data upon which it is based (round robin/test monitoring and dates) and
- Details of the reference oils/fuels (CEC reference number, batch number and brief description of product)

The statement must include the following summary statistics (as defined in section 2 of Procedure 4) for each measured parameter and reference oil/fuel:

- Mean
- Repeatability (*r*),
- Reproducibility (*R*)
- Number of test results in data set
- Number of contributing laboratories
- Number of outliers excluded

Precision statements should only quote general repeatability and reproducibility figures, or general equations relating *r* and *R* to performance level, if an adequate range of samples has been tested in a round robin, spanning the population of fluids falling within the scope of the test method. See section 5 of Procedure 1 for further details.

The precision statement must state the value(s), or range of values, of the measured parameter for which it is valid. It should also mention any limitations on the type of fluid to which it relates (e.g. is it valid for base stocks and/or formulated oils, fuels containing additives, oxygenated or bio components, metals, etc?).

Estimates of the repeatability *r* and reproducibility *R* are themselves subject to varying levels of uncertainty (see Table 1 in Appendix A of this Procedure and also Appendix B of Procedure 1). High levels of uncertainty in *r* and *R* are to be expected when these are calculated from small data sets. The precision statement must include appropriate warnings in such circumstances.

The definition of repeatability may need to elucidated in the precision statement to reflect how the data was collected and what changes in conditions (operators, equipment, materials, ambient, …) might have taken place between tests. For example, in round robins, repeat measurements on the same reference products are normally collected within a short time of one another while in test monitoring, repeat measurements are collected some time apart. See section 9 of Procedure 1 and section 2.5 of Procedure 2 for further discussion.

Optionally groups may include measures of accuracy (width of 95% confidence intervals) and discrimination (least significant differences at 95% confidence).

New test methods will be required to meet repeatability and/or reproducibility targets at each stage of the test development process set in accordance with sections 2 to 5 of this Procedure.

The precision statement must be reviewed after each round robin. The revised precision statement may incorporate data from previous studies if there has been no substantial change in precision (or severity) and if it is considered that including earlier data would give improved estimates or widen the range of applicability.

A precision statement review should also take place each time test monitoring data is analysed. Changes should be considered at this stage if there has been a substantial change in precision (or severity) or if, again, it is considered that including the new data would give improved estimates or widen the range of applicability. Care needs to be taken, however, as test monitoring data is not as well structured/balanced as round robin data and yields less reliable estimates of short-term repeatability.

After a precision statement review, the options are (a) to leave the statement as it is; (b) to update the statement using the new data in conjunction with some or all of the old data; (c) to revise the statement using only the new data. Option (a) might be chosen if there is little evidence of change in precision (or severity) and/or if the current data analysis is based on a small dataset or one which is not considered to be robust. Option (b) might be considered if the current statement is based on a relatively small data set, and if there is no evidence of major change in the new data. Option (c) would be chosen if there is clear evidence of a major change in precision (or severity) and this is based on a substantial data set. Further issues which should be considered are (a) how many laboratories and how much data the analysis is based upon and (b) whether the targets and limits are based on round robin or test monitoring data.

If there is evidence of appreciable changes in severity, and/or precision, then this needs to be reported to the working group and noted in the progress report.

Appendix A: Setting guidelines and targets for repeatability and discrimination at phase 1 of the Test Development stage

At the end of the Test Development stage, to be fit for use for the purpose(s) defined by the Management Board, a new CEC test should achieve appropriate statistical precision and discrimination levels. This Appendix provides guidelines to assist the Test Development Groups in developing guidelines and targets for repeatability and discrimination in the initial single laboratory phase 1. These targets are then sent to the Management Board for approval. Appendix C describes how repeatability and reproducibility targets should be set for phase 2 where a full round robin takes place across multiple laboratories.

When setting repeatability targets for phase 1, the TDG and Management Board will need to consider

- the uses to which the method will be put
- the test parameters for which guidelines/targets are to be set

Targets will normally be set for all primary parameters as defined in Section 1.

Working groups might also set repeatability targets for secondary parameters to ensure that they are fit for their intended purpose. Safety limits might also be set for "no harm" parameters.

If data or other relevant experience (e.g. from similar tests on other engines) is already available for the new test method, a target r_{target} or guidelines may be set for the repeatability r itself. If little or no data or experience is available, targets or guidelines may be based on the ability of the test to discriminate between particular test fluids.

Uses to which the method will be put

A test method that has been developed in a single laboratory may be considered fit for either of the following purposes:

- Comparing two fluids
- Checking conformance with relative specifications

once appropriate repeatability and discrimination levels have been demonstrated. The method may be only be considered fit for the above purposes when the tests are conducted at the test development laboratory. The method may not be considered fit for use at other laboratories until such laboratories have demonstrated satisfactory repeatability and comparability with the test development and other participating laboratories.

The method may not be considered fit for

- Estimation of the true value of the test parameter
- Checking conformance with absolute specifications

until the test has been rolled out to several laboratories and its reproducibility has been established.

Comparing two fluids

The ability of a test method to discriminate between two fluids depends on how different the fuels or lubricants actually are (see Procedure 4 section 7). Significant differences are much more likely to be observed for pairs of fluids with very different performance levels than for pairs which differ only a little.

> If two samples are tested at the same laboratory, then the measured difference will be statistically significant at the 95% confidence level (2-sided test) if it exceeds the repeatability *r*.
>
> There will be a 50% chance or greater of measuring such a difference if the (unknown) true value of the difference exceeds *r*
>
> And a 95% chance or greater of measuring such a difference if the true value of the difference exceeds 1.84*r*
>
> *(Note: better discrimination can be achieved by performing repeat tests on the two samples at the same laboratory; see Procedure 4 for discrimination calculations for other experimental designs)*

To calculate a repeatability target, you could consider the sorts of fluids you would like to be able to differentiate between. Suppose you wish to be able to distinguish two fluids differing in performance by *D* test units or more. Then you could set

$$r_{target} = D$$

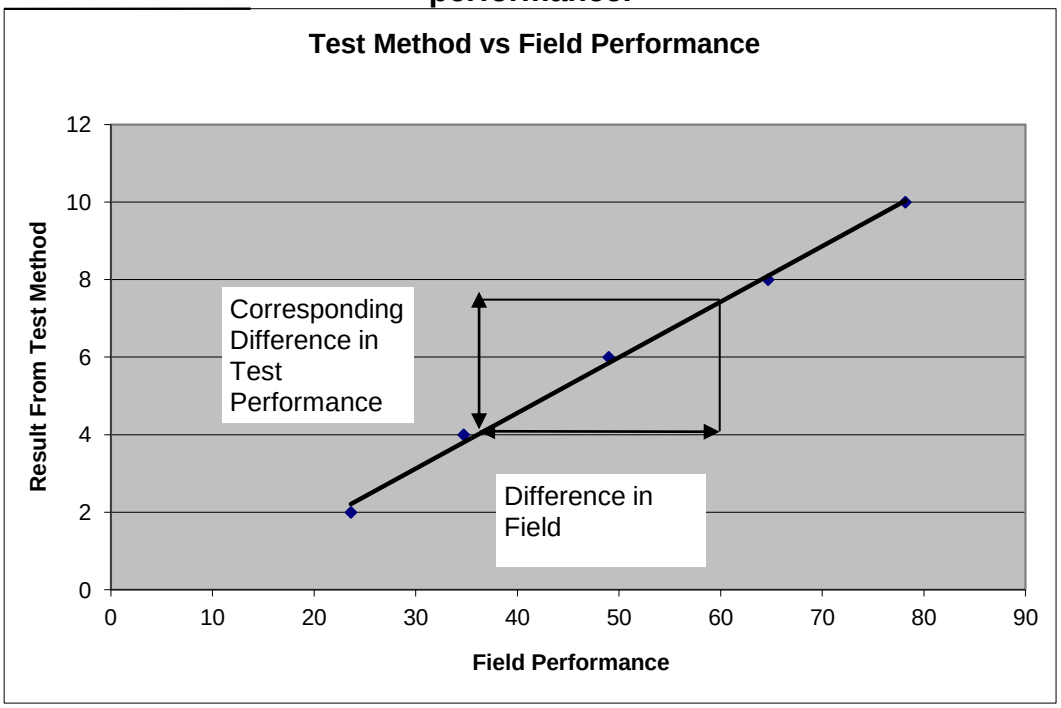
if you require a 50% chance of seeing a significant difference, or

$$r_{target} = D / 1.84$$

if you require a 95% chance of seeing a significant difference between two test fluids differing in performance by exactly *D* test units.

When selecting *D*, you could take into account differences of practical importance in the field as illustrated in Figure 1.

Figure 1. Typical plot showing correlation between test results and field performance.



Quite often, although no test data will be available for the new method, it may be possible to select (for example) "good", "marginal" and "poor" test fluids using scientific judgement and experience. Guidelines may then be set based on the method's ability to discriminate between these fluids based on the ratio of the repeatability *r* to the performance difference between two such fluids Δ measured in the test development programme.

If the ratio

$r / \Delta < 1$ (equivalent to $DP(r) / \Delta < 1.84$)

then there is a 50% or greater chance of seeing a significant difference between two such fluids in future. If the ratio

$r / \Delta < 1 / 1.84 = 0.54$ (equivalent to $DP(r) / \Delta < 1$)

then there is a 95% or greater chance of seeing such a difference.

Checking conformance with relative specifications

Sometimes the prime purpose of a test method will be to check conformance with relative specifications that is to compare a candidate fluid with a reference fluid. In some specifications, the candidate will simply need to obtain a result better than the reference fluid to pass – in others, the candidate will need to better the reference by more than a specified margin *M*.

To check conformance with relative specifications, independent tests should be conducted on the candidate and reference fluids under the same conditions (same engine or stand, same laboratory, same operator) within a short time of one another.

If there is a delay between the candidate and reference tests, then the comparison may be contaminated by drift in test conditions.

The probability that a candidate fluid will pass the test depends on how different the candidate and reference oils or fuels actually are. Significant differences are much more likely to be observed for pairs of fluids with very different performance levels than for pairs which differ only a little.

A candidate has a 50% probability of passing the test if its true performance is the same as that of the reference. It has a 95% probability of passing the test if it is better than the reference by 0.84*r*. Therefore one way to set a repeatability target is to think of the sorts of candidate fluids which one would want to pass the test on most occasions. Suppose these are candidates which better the reference by X test result units or more. If the repeatability target is set at

$$r_{target} = X / 0.84$$

then any such candidate will have at least a 95% chance of passing the specification.

If the candidate has to be better than the reference by a margin *M* to meet the specification, then the repeatability target should be set at

$$r = (X - M) / 0.84$$

Example

If a candidate fuel needs to better a reference fuel by 5 test units to meet a specification, and if the user requires a candidate which is 10 units better than the reference to have a high 95% probability of passing the specification, then r_{target} should be set at

$$r_{target} = (10 - 5) / 0.84 = 6.0 \text{ test units}$$

Estimation of the true value of the test parameter & checking conformance with absolute specifications

If the prime purpose of the test is to estimate the true value of the test parameter or to check conformance with absolute specifications, then the method will need to have good reproducibility. Reproducibility cannot be estimated at the initial single-laboratory phase 1 of the test development process. Nevertheless, the reproducibility requirements can be taken into account when setting repeatability guidelines and targets.

Assessing the reproducibility required in a new test can be difficult if little is known about test severity and how fuels and lubricants might eventually perform. However in cases where the new test is a replacement for an existing test, or where the result of the new test is a simply understood quantity such as fuel economy improvement or viscosity, it may be possible to set a reproducibility target R_{target} using the guidelines in Appendix C. If a reproducibility target R_{target} can indeed be set, then the repeatability target r_{target} could be set using the formula

$$r_{target} = k \times R_{target}$$

where the multiplier *k* is a number between 0 and 1 chosen to reflect the typical repeatability / reproducibility ratio for tests similar to the one under development. The multiplier *k* may thus be established by looking at the relative repeatability / reproducibility ratios for approved tests of the same type.

The reproducibility *R* of a test method must by definition be greater than or equal to the repeatability *r*. The ratio $k = r / R$ will typically vary between 0.3 for tests where there is a substantial degree of laboratory-to-laboratory variability (e.g. octane or cetane ratings) to 0.9 for tests where laboratory-to-laboratory variability is very small (e.g. tests that involve a measurement relative to an internal reference such as fuel economy improvement tests).

The eventual purpose of a new test may be to check conformance with absolute specifications. However it cannot be used for this purpose until its reproducibility *R* has been established. As a short term measure, a relative specification might be set at the end of the test development phase (or any pilot inter-laboratory programme) under which tests are conducted on a candidate fluid and reference fluid at the test development (or other accredited) laboratory only. The candidate meets the specification if it achieves a better result than the reference. The capability of the test to check conformance with the relative specification will be equivalent to its capability to check conformance with the absolute specification if $r = 0.71R$. Therefore r_{target} should be set to

$$r_{target} = 0.71 \times R_{target}$$

if R_{target} can be established by the methods in Appendix C.

Warnings

1. Cost constraints may limit the number of tests that can be conducted in phase 1 of the test development process to establish repeatability and discrimination. As a consequence the calculated repeatability may itself be subject to a substantial degree of uncertainty.

Table 1 gives 95% confidence limits for the true repeatability at the test development laboratory as a function of the measured repeatability *r*

Table 1. 95% confidence limits for the true repeatability as a function of a measured value *r* and its associated degrees of freedom.

d.f.	95% confidence limits
1	0.446*r* - 31.910*r*
2	0.521*r* - 6.285*r*
3	0.566*r* - 3.729*r*
4	0.599*r* - 2.874*r*
5	0.624*r* - 2.453*r*
6	0.644*r* - 2.202*r*
7	0.661*r* - 2.035*r*
8	0.675*r* - 1.916*r*
9	0.688*r* - 1.826*r*
10	0.699*r* - 1.755*r*
15	0.739*r* - 1.548*r*
20	0.765*r* - 1.444*r*
25	0.784*r* - 1.380*r*
30	0.799*r* - 1.337*r*

The multipliers in this table may also be used to calculate 95% confidence limits for the true reproducibility as a function of a measured value *R*

where the number of degrees of freedom is

d.f. = (no. of samples) × (no. of repeat tests on each sample – 1)

Thus if 3 tests (say) are conducted on each of 2 samples in the phase 1 programme, then there will be 4 d.f. and so the true repeatability could be anywhere between 0.6× and 2.9× the value actually measured.

As a result of this imprecision there will, in many cases, be uncertainty as to whether a test does in reality meet a repeatability target or not. In such cases, the Management Board, technical experts and the SDG member will need to make a careful assessment of the performance of the test taking into account both statistical analysis and engineering judgement. Individual test results are likely to have large influences on estimates of repeatability and the technical experts will need to make a thorough examination of any major deviations in test results.

2. Repeatability is paradoxically rather harder to define than reproducibility. Reproducibility measures the closeness of agreement between test results conducted on identical samples at different laboratories. Thus the laboratory, the test engine or stand, and the operators are all different.

Repeatability measures the closeness of agreement between test results conducted on identical samples at the same laboratory. This will depend on the elapsed time between tests and on whether different test engines or stands, or different operators can be used. CEC Procedure 4 defines repeatability as

- 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, in a short interval of time, and under the same test conditions (same operator, same apparatus, same laboratory), may be expected to lie with a probability of 95%

It is recommended that in CEC round robins, the test fluid should be changed after every test so the repeat tests on a particular fluid are conducted within a short time of one another, but not back-to-back. Back-to-back tests on the same fluid are unlikely to be truly independent and will lead to estimates of repeatability which are artificially small and which give an over-optimistic assessment of the discriminating power of the test. See section 5 of Procedure 1 for further discussion.

3. The ability of a test to discriminate between fuels and to check conformance with relative specifications may deteriorate if there is a delay between the two tests.

Appendix B: Accreditation of new laboratories in the early stages of test introduction

The first tests outside the test development laboratory will serve a dual purpose. They will help assess both the portability of the test and also the ability of the participating laboratories to obtain similar results to the test development laboratory.

In the early stages of test dissemination and in the first round robin, it is recommended that each participating laboratory should perform at least two tests on at least two test fluids spanning the performance range of interest with regard to the prime test purpose. Ideally considerably more tests should be conducted.

During the early stages of test introduction, new laboratories may be considered to meet CEC's quality requirements, as given in Guideline 18, if they achieve similar test result levels and similar repeatability and discrimination to the test development laboratory. Guidelines for determining whether the results and precision are acceptable are given below.

1) Consistency with the Test Development Laboratory

A guideline may be issued requiring some or all of the following:

- All test results to fall within $\pm k_1$ SD of the average result obtained by the test development laboratory on that particular test fluid
- The average difference in test results between a new laboratory and the test development laboratory to be less than $\pm k_2$ SD.
- Let Δ be the average difference between 2 reference products. The average difference in Δ between a new laboratory and the test development laboratory to be less than $\pm k_3$ SD.
- The root mean square deviation between a new laboratory's result and the development laboratory's mean result for that sample to be less than k_4 SD

The appropriate standard deviations SD and multipliers k_i (i = 1 - 4) should be determined in consultation with a SDG member. Larger multipliers might be used for tests with multiple parameters to compensate for the increased risk of at least one value drifting outside the above limits due to random variation.

2) Consistency with Other Laboratories participating in the round robin.

A guideline may be issued requiring some or all of the following, where RSD is the estimated reproduciblity SD:

- All test results to fall within $\pm k_1$ RSD of the average result obtained by all laboratories on that particular test fluid
- The average difference in test results between a new laboratory and the industry to be less than $\pm k_2$ RSD.
- Let Δ be the average difference between 2 reference products. The average difference in Δ between a new laboratory and the industry mean to be less than $\pm k_3$ SD.
- The root mean square deviation between a new laboratory's result and the industry mean result for that sample to be less than k_4 SD

3) Consistency requirements for laboratories introducing the test after the round robin

Once the first round robin is complete and the test is accepted, Procedure 2 section 1.14 may be used to establish rules for the addition of new installations.

Appendix C: Setting the repeatability and reproducibility targets for a full round robin

Once the phase 1 (single laboratory) test development is approved by the Management Board, and any pilot inter-laboratory programmes are complete, a round robin shall be conducted at a wider population of laboratories to determine the reproducibility of each test method parameter and to confirm its repeatability at other locations (phase 2).

A minimum of 5 laboratories is required for the full round robin. In situations where there are less than 5 laboratories with equipment installed, it will not be possible to estimate the reproducibility of the test method to an acceptable degree of precision. In such circumstances, the approach will be to compare test result levels, precision and discrimination at each new laboratory against those at the test development laboratory (see Appendix B).

If the prime purpose of the test is

- Comparing two fluids or
- Checking conformance with relative specifications

then the relevant test parameter must demonstrate appropriate repeatability and discrimination.

If the prime purpose of the test is

- Estimation of the true value of the test parameter
- Checking conformance with absolute specifications

then the relevant test parameter must demonstrate appropriate reproducibility.

Working groups might also set reproducibility targets for secondary parameters to ensure that they are fit for their intended purpose. Safety limits might also be set for "no harm" parameters which the various reference fluids, and most candidates, would be expected to meet on a consistent basis (see Procedure 4 Section 8). However there may be no requirement to discriminate between reference fluids if these are all expected to pass a specification or safety limit.

To meet CEC quality standards (see Guideline 18), a laboratory must demonstrate that it can achieve similar test results and repeatability to its peers in accordance with the guidelines in Appendix B.

To be accepted as fit for the purpose defined by the Management Board, a CEC test parameter must achieve a repeatability r and/or a reproducibility R less than or equal to a repeatability target r_{target} or reproducibility target R_{target} in a round robin where r_{target} and/or R_{target} are agreed by the Management Board and Test Development Group, following the process described in Section 4. This Appendix gives guidelines for setting the repeatability and/or reproducibility targets.

When developing a repeatability and/or reproducibility target, the participants will need to consider

- the uses to which the method will be put
- the test parameters for which guidelines/targets are to be set
- criticality of the particular test parameter
- differences of practical importance in the field

The formulae at the end of this Appendix, and the examples in Figures 2 and 3, may then be used to help decide on suitable values for r_{target} and/or R_{target}.

After meeting the repeatability and reproducibility targets for a particular parameter, no further improvements its precision need be sought.

Uses to which the method will be put

A test method may be considered fit for use for either of the following purposes:

- Comparing two fluids
- Checking conformance with relative specifications

at a particular laboratory once appropriate repeatability levels and discrimination have been demonstrated for the parameter of interest at that laboratory.

The method may only be considered fit for

- Estimation of the true value of the test parameter
- Checking conformance with absolute specifications

once the test has been rolled out to several laboratories and appropriate reproducibility has been established for the parameter of interest.

Formulae for evaluating test performance are given at the end of this Appendix.

The test parameters for which guidelines/targets are to be set

The Management Board will decide the test parameters for which guidelines/targets are to be set paying due regard to the prime purpose of the test.

Criticality of the particular test parameter

When deciding on the level of test method performance you would like to see, you need to consider not only how the method is going to be used but also how critical each measured parameter is likely to be. For example, a temperature difference of 0.1 C may be of critical importance to a physician, while a motorist is happy with a dial going from C to H.

As a general guideline, reproducibility targets should be more stringent, i.e. smaller, for critical parameters than for non-critical ones.

Differences of practical importance in the field

Most CEC methods are designed to correlate with some aspect of field performance. If you have field performance data available, you should plot test results against field performance measurements, as shown in Figure 1 from Appendix A for example.

This will enable you to translate a difference in field performance, which is of interest to you, into the corresponding difference in test performance. If you do not have quantified field performance data available, but can still categorise oils or fuels as "good", "marginal" or "poor" by some means, then you can use these labels on the x-axis.

It will not be possible to correlate field performance with test results until the test method is in its final form. Therefore final repeatability and reproducibility targets should not be set until the method has been fully developed and no further major changes, e.g. in severity, are envisaged.

When considering a new method, you should think about what level of measurement accuracy or performance differentiation is required, always bearing in mind the link between test results and field performance. For example, potential users of the new test method illustrated in Figure 1 might need a good chance of discriminating between two oils differing by 25 units in field performance and thus by 3 units in test performance. In a different application area, a typical customer might need a test that is capable of measuring viscosity to within ±3%.

Formulae for evaluating test performance

We now give formulae for evaluating various aspects of the performance of a test method as a function of the repeatability *r* and reproducibility *R* for each measured parameter. These are followed by examples showing how the calculations might be performed on a spreadsheet and how spreadsheets might be used in setting reproducibility targets.

It will be assumed, for simplicity, that only one test parameter is of interest. It is difficult to give generic formulae for comparisons or specifications involving two or more test parameters as the various probabilities will depend on the degree of correlation between those parameters. The WG Chairman should seek advice from the SDG Liaison Officer in such cases. Multiple parameters are discussed further in Procedure 4 section 8.

Repeatability

> There is a 95% chance that two measurements on the same sample taken at the same laboratory under the same conditions (same operator, same apparatus, same laboratory, short interval of time) will lie within +/- *r* of one another

Reproducibility

> There is a 95% chance that two measurements on the same sample taken at different laboratories will lie within +/- *R* of one another

Estimation of true value

> If a single measurement is made on a particular sample, then a 95% confidence interval for the true value will be
>
> measured value +/- 0.71*R*

Discrimination

If two samples are tested at the same laboratory, then the measured difference will be statistically significant at the 95% confidence level (2-sided test) if it exceeds the repeatability *r*

If two samples are tested at different laboratories (a poor experimental design), then the measured difference will be statistically significant at the 95% confidence level (2-sided test) if it exceeds the reproducibility *R*

There will be a 50% chance or greater of measuring such a difference if the (unknown) true value of the difference exceeds *r* (tests at same laboratory) or *R* (tests at different laboratories)

And a 95% chance or greater of measuring such a difference if the true value of the difference exceeds 1.84*r* or 1.84*R* respectively.

(Note: better discrimination can be achieved by testing the two samples more than once at the same laboratory; see Procedure 4 for discrimination calculations for other experimental designs)

Margins

Absolute specifications

To be 95% confident that his product meets a lower specification limit A_L, a supplier needs to obtain a test result *X*, which is greater than A_L by a margin of 0.59*R* or more

By the same token, if the true performance of a product is 0.59*R* greater than A_L then the producer has at least a 95% chance of obtaining a test result greater than A_L

Similar margins apply in the opposite direction for upper specification limits A_U

Relative specifications

To be 95% confident that a candidate fluid is really better than a reference fluid, a supplier needs to obtain a candidate test result which is better than the reference test result by a margin of 0.84*r* or more.

By the same token, if the true performance of the candidate is 0.84*r* better than that of the reference then the producer has at least a 95% chance of obtaining a better test result on the candidate fluid.

Absolute specification setting

A two-sided specification (which can be failed in both directions) is only allowable under International Standard ISO 4259 if the reproducibility *R* is no more than a quarter of the specification width.

A one sided specification (which can only be failed in one direction) is only allowable under ISO 4259 if the reproducibility is no more than half the specification width. For example, if the specification requires that a length measurement be less than 10 microns, the specification width is 10 microns since negative lengths are impossible. The reproducibility thus needs to be less than 5 microns.

More extensive formulae (including those needed for different confidence levels) may be found in Procedure 4.

Figure 2 shows how you might perform these calculations on a spreadsheet once the repeatability *r* and/or reproducibility *R* have been estimated after a particular phase in the test development and surveillance cycle.

Figure 2. Use of spreadsheets to assess test performance. The table shows how a test method would perform if its repeatability *r* was 5.0 and its reproducibility *R* was 8.0 in a round robin.

Reproducibility *R* = 8.00	**Formula**	**Performance**
Reproducibility	*R*	8.0
Estimation of true value (95% confidence limits) ±	0.71*R*	5.7
Absolute specification margin (95% prob)	0.59*R*	4.7
Smallest absolute specification range (1-sided)	2*R*	16.0
Smallest absolute specification range (2-sided)	4*R*	32.0

Repeatability *r* = 5.00	**Formula**	**Performance**
Repeatability	*r*	5.0
Discrimination (50% prob.)	1*r*	5.0
Discrimination (95% prob.)	1.84*r*	9.2
Relative specification margin (95% prob)	0.84*r*	4.2

The formulae above can be inverted to derive repeatability and reproducibility targets at the start of each test development phase. This is illustrated in Figure 3.

Figure 3. Use of spreadsheets to determine r_{target} and R_{target}. First choose the desired level of performance with respect to the prime purpose of the test. For example, this might be (a) to check conformance with an absolute specification of X < 10 test units or (b) to be able to discriminate between two fluids 5 test units apart with a high degree of confidence. Then set the repeatability target r_{target} or reproducibility target R_{target} to achieve this level of performance.

Reproducibility target setting	**Desired performance *P***	**Formula**	***R*$_{target}$**
Reproducibility		*P*	
Estimation of true value ±		*P*/0.71	
Margin to pass absolute specification (95% prob)		*P*/0.59	
Absolute specification setting (1-sided)	10	*P*/2	5.0
Absolute specification setting (2-sided)		*P*/4	

Repeatability target setting	**Desired performance *P***	**Formula**	***r*$_{target}$**
Repeatability		*P*	
Discrimination (50% prob.)		*P*	
Discrimination (95% prob.)	5	*P*/1.84	2.7
Margin to pass relative specification (1-sided) (95% prob)		*P*/0.84	

If more than one aspect of performance is of importance, then calculate the repeatability and/or reproducibility targets required for each aspect and select the smallest value. Thus if a reproducibility target of 5.0 is needed for specification setting and a reproducibility target of 3.5 is needed for estimating the true value, then the reproducibility target should be set at 3.5.

Test methods where with the variability depends on the mean

If the variability of the test parameter(s) changes with the mean test result level, the TDG Chairman should seek advice from the SDG Liaison Officer. In the simplest case it may be possible to express the repeatability and reproducibility targets as percentages of the mean. In more complex cases the precision targets would be specified at a fixed mean level. An alternative is to set precision targets for a particular reference fluid, which will usually be of borderline performance.