Procedure 1

Round Robins

Procedure 1

Round Robins

This procedure replaces former Procedure 1 - Conduct of Precision Testing Programmes and Procedure 2 - Collection, Transfer and Reporting of Reference Test Data.

1. What is a Round Robin?

A “round robin” is a test programme in which a number of laboratories test identical samples of a number of test materials in order (primarily) to determine the precision (repeatability and reproducibility) of each reported parameter in a test method.

Example

Table 1 shows data from a round robin to determine the precision of a test method that measures the kinematic viscosity of used engine oils. Four used oil samples A, B, C and D were tested in duplicate at 12 laboratories.

After the rejection of several abnormal sets of data points (shaded in yellow) as “outliers” using the methods detailed in International Standard ISO 5725-2 [1] and Appendix A of this procedure, the repeatability *r* and reproducibility *R*, as defined in Procedure 4, were estimated to be

$$r = 0.0314 \times y - 0.275;\ R = 0.243 \times y - 5.43$$

where *y* denotes the true viscosity of the oil under test.

(Note: this round robin was conducted in 2002 and only tested 4 samples; the current procedure requires a minimum of 5 samples in order to determine *r* and *R* as functions of *y*)

2. Purpose

Round robin programmes are conducted in order to understand and quantify the variation in each reported parameter in a test method. They may also be used to measure the performances of particular fluids or fluid batches and/or to investigate the severity[1] of a test method.

[1] Severity: In CEC, the term “severity” is used rather loosely to express the position(s) of the mean(s) of set(s) of test results on a particular sample or samples. A test is described as becoming more “severe” if changes in mean test results occur that are indicative of worsening test fluid performance and “mild” if the changes are indicative of better performance. Laboratories can also be described as “severe” or “mild” relative to their peers if there are systematic differences in their test results.

Table 1. Data from a round robin to determine the precision of a test method that measures the kinematic viscosity of used engine oils (SG-L-083).

KV100, cSt	A	B	C	D
Lab 2	20.71	83.32	35.68	82.10
Lab 2	20.96	84.05	35.81	82.04
Lab 3	20.34	80.80	34.97	81.58
Lab 3	20.26	81.19	34.98	81.66
Lab 4	20.53	80.37	35.71	80.93
Lab 4	20.42	81.85	35.18	82.81
Lab 5	20.03	72.36	34.81	69.57
Lab 5	19.94	73.30	34.77	70.63
Lab 6	2.60	2.50	2.10	2.70
Lab 6	2.50	2.40	2.10	2.70
Lab 7	20.55	82.34	36.71	81.31
Lab 7	21.16	83.40	37.50	81.31
Lab 8	20.40	75.53	35.13	73.87
Lab 8	20.31	74.84	34.68	71.50
Lab 9	20.19	75.10	34.82	72.70
Lab 9	20.35	78.61	34.93	73.04
Lab 10	20.75	77.67	35.73	76.04
Lab 10	20.76	77.70	35.74	76.03
Lab 11	20.31	74.25	34.38	75.80
Lab 11	20.27	74.50	34.46	76.37
Lab 12	20.51	74.13	34.77	72.57
Lab 12	20.75	75.48	35.10	71.64
Lab 13	20.46	75.22	32.23	65.65
Lab 13	20.39	79.47	33.04	76.38

For new tests, the key objective is to establish the precision and produce precision statements. Round robins can also be used to monitor the precision of existing tests and update their precision statements if appropriate (see Procedure 3).

The precision statistics (repeatability and reproducibility) determined from the round robin can then be used

… to evaluate the method's suitability for measuring the performance of products

… to check the achievement of repeatability and reproducibility targets

(see Procedure 3 for details).

The statistical analysis may also be used:

… to compare laboratories in terms of their precision (repeatability), severity and ability to rank fluids consistently with their peers

… to measure the performance of new reference samples/batches (e.g. to set test monitoring targets and control limits; see Procedure 2) or new types of fluid

… to check for severity and precision changes (perhaps after hardware or other method changes)

… to study the impact of other factors on the precision or severity

Round robin studies may also be used to try and understand the sources of variability in test results in order to improve a method's precision. Laboratories might therefore be requested to supply additional information about their installation, or to gather additional data during each test (e.g. on ambient or engine-running conditions) in the round robin.

Most tests consist of producing a phenomenon, e.g. wear, and then measuring it. For investigative purposes, the measurement aspect can be studied independently, for example in rating workshops. Round robins can also be carried out solely on the measurement part of a test procedure. For example, the increase in viscosity as the soot content of a fresh oil builds to 6% is one of the key outputs from the PSA-DV4 engine test (SG-L-093). The viscosity increase is measured using the measurement method studied in Table 1 (SG-L-083).

The precise objectives of round robin programmes must agreed at a Working Group meeting prior to the commencement of the study and shall be recorded in the minutes.

3. When Should a Round Robin be Run?

When a new test method is being developed, the first round robin will normally be conducted at the end of the single-laboratory test development phase when the method is rolled out to other laboratories.

Subsequent round robins should then be conducted on a regular basis, typically annually, until the method is stable unless otherwise agreed by the Working Group and Management Board.

All laboratories in the Working Group must participate in round robins.

Once the method is considered stable and no major changes in severity or precision are being observed, the group may reduce the frequency of round robins, and use test monitoring data from the CEC-TMS or ATC-ERC reference database, as appropriate, to monitor severity and precision (see Procedure 2). However migration to test monitoring may prove impractical when reference fluid batches have too short a shelf life to establish targets and thereafter collect reasonably lengthy series of repeated measurements.

If major changes are made to the test method, a round robin will be needed to check for changes in precision and/or severity.

Mini round-robin programmes may also be required to determine the properties of new reference fluid batches. These might involve just one sample and limited numbers of evaluations, the size of the round robin being particularly constrained in expensive engine tests.

Pilot studies

When the single-laboratory test development stage is complete, the working group may conduct pilot inter-laboratory programme(s) 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

The working group may also decide to conduct a mini round robin (one test on one or two samples per laboratory) in order to obtain a preliminary estimate of reproducibility across a wider population of laboratories.

The results from a pilot programme or mini round robin may be considered as forming part of a larger round robin if the study is subsequently extended to further laboratories testing the same sample(s) within a reasonable period of time.

4. Responsibilities

The WG (Working Group) Chairman has overall responsibility for organising the round robin. The chairman must ensure that the WG members set clear objectives and time scales.

All laboratory members of the WG shall take part in the round robin and provide test results within time scale agreed.

The Statistical Development Group Liaison Officer (SDG LO) shall provide help in designing the round robin to meet the Working Group's objectives.

The WG Chairman may appoint a Working Group Database Administrator (WG DBA) to collate round robin and test monitoring results. The SDG LO and the WG DBA will agree the format in which the data will be transferred (see section 7 below).

The SDG LO will perform the statistical analysis of the final results and present these to the WG.

The SDG LO will provide other analyses and advice to the WG that may be useful for improving their understanding of the test.

5. Design

Number of laboratories

All laboratory members of the WG shall take part in round robins. At least five laboratories/stands are required in order to obtain a reasonably precise estimate of reproducibility, but it is preferable to have more. The number of laboratories/stands participating in a round robin shall be stated in any precision statement based on the results.

If the total number of laboratories in the WG is less than five, then the round robin can and should still take place. In such circumstances, statistical advice should be sought on how any general reproducibility figure should be interpreted, and appropriate caveats must be given in the precision statement.

Number of samples

The number of samples shall be sufficient to span the population of fluids falling within the scope of the test method, and it should also cover the likely ranges of each reported parameter. Issues to be considered might include, for example, oil viscosity grade, fuel composition, presence/absence of additives, etc. Methods should correlate with field performance.

If any variation in precision with performance level has been observed in previous programmes, or is expected from experience with similar tests or from engineering judgement, then at least five samples need to be tested if the aim is to express repeatability *r* and reproducibility *R* as functions of level. The number of samples required will be greater when the relationships between *r* and *R* and performance level are nonlinear.

It is recognised, however, that many CEC tests, particularly engine tests, are expensive and the cost of testing five samples at every laboratory may be prohibitive. In such circumstances, a smaller number of samples may be tested; a minimum of two is required. Fewer samples might also be tested in periodic round robins on stable methods, or when linear relationships between precision and level, or suitable variance stabilising transformations[2] have been found in previous exercises.

[2] International Standard ISO 4259 [2] Annex E details a number of such transformations, e.g. log or arcsin.

Table 2. 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*

When fewer than five samples are tested, however, it may subsequently prove impossible to infer the values of the repeatability *r* and reproducibility *R* for other samples at different performance levels. In such circumstances, precision statements should simply quote the values of *r* and *R* for the samples tested.

When precision depends on level, the working group may base its repeatability and reproducibility targets on one particular sample in the round robin (see Procedure 3). If the method is used in a specification, the chosen sample will typically be of borderline performance.

Number of repeats

In order to estimate repeatability, each sample will normally need to be tested twice at each laboratory. Further repeats may be required to obtain a reasonably precise estimate of repeatability if the number of participating laboratories is small. The total d.f. (degrees of freedom) for repeatability on a particular sample is

Repeatability d.f. for single sample
= Total number of tests on that sample – No. of laboratories testing that sample

Table 2 may be used to determine 95% confidence limits for the true repeatability and reproducibility of a test method as a function of their degrees

Table 3. Measurements of inlet valve cleanliness rated on a 0-10 scale in a SG-F-005 round robin.

	Sample A		Sample B	
Laboratory	Test 1	Test 2	Test 1	Test 2
1	7.55	7.70		
2	7.82	7.26		
3	8.55	7.68		
4	7.50	7.73		
6			9.44	9.48
8			9.78	9.13
9			9.75	9.65
10			9.51	9.62
11	7.78		9.53	
12	8.15		9.53	
14	7.66		9.93	
15	7.98		9.50	
7	7.33		9.32	
18	9.38		9.10	

of freedom[3]. For example, if 5 laboratories test 2 samples in duplicate then there will be 5 repeatability d.f. per sample and the true repeatability will lie between 0.624× and 2.453× the measured repeatability with 95% confidence. This may be deemed too imprecise in which case more repeats will need to be conducted at each laboratory.

If testing is very expensive, working groups can consider alternative smaller designs in which laboratories test some samples once and others twice.

Example

Table 3 shows data from a round robin to determine the precision of engine-test ratings (on a 0-10 scale). Four laboratories tested sample A twice, four laboratories tested sample B twice and six laboratories performed one test on sample A and one test on sample B. This design yields 9 or more d.f. per sample for estimating the reproducibility R, but only 4 d.f. per sample for estimating the repeatability r.

The working group and SDG LO will need to balance the accuracy of the precision estimates produced from the round robin against the cost of testing when deciding on the number of repeats and the possible use of smaller designs such as that employed in Table 3; such designs also may lead to

[3] The reproducibility d.f. depend on the measured data (see Appendix B) and so cannot be determined until the round robin is complete. However, as a rule of thumb, the reproducibility d.f. for a particular sample will generally be slightly greater than the number of laboratories measuring that sample minus one.

problems in the analysis, particularly when the number of d.f. for repeatability is small (see Appendix A, section 7.4.5.4).

In most round robins, the estimation of reproducibility will take precedence over that of repeatability. This means that usually, all labs should test all samples. An exception is studies where the prime purpose of the test method is to compare the performance of different fluids at the same laboratory, for example to check conformance with “relative to reference” specifications (see Procedure 3).

Test order

The full set of round robin tests at any particular laboratory should ideally be completed as a continuous programme in as short a time as is practicable.

The samples should be tested in random order at each laboratory, with a change of sample between each pair of successive tests.

Repeat tests on the same sample at the same laboratory must be conducted independently as if they were tests on different materials. Two (or more) tests on the same sample should 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. If in situations where operators know they are performing repeat tests, it is feared that previous results may influence subsequent test results, then it may be necessary to blind code samples in such a way that operators will not know which are the replicates.

(Note: The recommended test order differs from the practice in ISO 5725 [1] and ISO 4259 [2]. CEC procedures require a change of test fluid on every test so that the repeatability calculated from the round robin results provides an appropriate error estimate when comparing different fluids)

Randomised block designs may be used to determine the test order. For example, the following test order may be used at a laboratory which is testing four samples A, B, C and D in duplicate:

A	C	B	D	C	A	D	B

Each of the four samples is tested once in randomised order; then each sample is tested again in a different randomised order. There must not be a serious break or change in test conditions or operators between blocks.

Different random orders should be used at each laboratory.

If a test is aborted or found to be invalid by the test laboratory during the round robin, then it should be repeated as soon as is practicable.

Timing

Round robin programmes measure the precision of a test and the condition of test fluid batches at a particular point in time. Therefore while different laboratories cannot be expected to conduct their tests on exactly the same day(s), it is important that the participating sites complete their tests within a defined time window. This will normally be of three months duration or less.

6. Conduct, sample handling and consumables

All tests must be conducted in strict accordance with the appropriate CEC test procedure and must be completed within the time frame specified. Any additional instructions from the working group related to the particular round robin must be followed.

All tests at a particular laboratory shall be conducted by the same operator(s) using the same equipment. If a laboratory has more than one stand/instrument then it may be asked to submit independent sets of test results from two or more of these installations. In such circumstances, the results from the various stands would be treated as if they came from different laboratories in the statistical analysis. Therefore it is reasonable and indeed desirable to assign different operators to different stands/instruments.

All the laboratories/stands must use identical batches of reference oil or fuel as directed by the working group chairman or appointee. Proper procedures must be developed for the manufacture, distribution, transport and storage of samples so that these remain “identical” and homogeneous at the time of use.

Samples that are expected to be stable may be accumulated, subdivided and distributed by the organizer or his appointee. Adequate quantities must be prepared to allow for errors, spillages etc. When samples are standard CEC reference fluids, these may be obtained directly from the supplier, but care must be taken to ensure that all laboratories use the same fluid batch.

Some test methods are intended to be used on less stable samples, for example fuels containing oxygenates or metals. In such circumstances, it may be prudent to mix components at the test laboratory just before testing. However some thought needs to be given to reproducibility estimates derived from such studies. Such estimates will exclude variations associated with sampling, transport and storage. These variance components might be thought of an integral part of laboratory-to-laboratory variability if the method is to be used to evaluate samples taken from the field.

Reproducibility estimates derived from round robins where samples are blended at the test laboratories may be artificially small and give a false impression of the accuracy of the test in the field. The precision statement should make it clear how and where round robin samples were blended in situations where sample stability might be an issue.

When test methods use consumable parts (e.g. pans or pistons), materials (e.g. seals) or fluids (e.g. lubricants in fuel tests, fuels in lubricant tests,

coolants), then it is possible (but not desirable) that these could come from different manufacturers or from different batches. Such consumables may vary from laboratory to laboratory and even, over the course of time, within a laboratory. In some round robin exercises, laboratories have been asked to use consumables from the same manufacturer/batch in order to reduce variability. However such controls can lead to reproducibility estimates which are artificially small. Such reproducibility values will underestimate future levels of variability if laboratories remain at liberty to use different manufacturers and batches. Any precision statement based on such round robins must mention any artificial controls on consumable parts, materials or fluids.

7. Data transfer

All round robin data must be transferred electronically to minimise the risk of data corruption. The WG DBA is responsible for collating the results.

Working Groups that use the CEC-TMS reference database for test monitoring, or who register results in the ATC-ERC database, should normally use the same database to collect round robin data. Round robin results will need to be coded to differentiate these from test monitoring results on the same reference samples. The WG DBA will need to liaise with the CEC-TMS or ATC-ERC database administrator to ensure that samples can be coded appropriately

Spreadsheets may provide a more appropriate data transfer medium if information is to be collected on additional parameters that are not included in the data dictionary, or if non-standard reference fluids are to be tested. These will be collated by the WG DBA. However laboratories must still, in addition, submit the results of those round robin tests conducted under normal conditions to the appropriate database in accordance with its regulations (see Procedure 2).

Working Groups not in the test monitoring system, for example TDGs (Test Development Groups) or new SGs (Surveillance Groups), must use other electronic means of collating data from round robins. Again spreadsheets are recommended.

If round robin data are stored in the CEC-TMS or ATC-ERC database, these should be extracted by the WG DBA when the round robin is complete and sent to the SDG LO for analysis. The WG DBA should also collate any comments on the validity of tests at the various laboratories. The SDG LO may extract the data directly if agreed by the WG DBA.

When spreadsheets are used to transfer data, the SDG LO and WG DBA shall design suitable data entry templates, typically blank spreadsheets, which will be distributed to participating laboratories by the WG DBA.

Data entry spreadsheets should be arranged so that data from different laboratories can be collated easily into a format suitable for analysis, by the SDG LO, using START or other statistical analysis programs (see section 8 below). Normally the spreadsheet will be arranged so that one row corresponds to one test and each column corresponds to a different variable. Space must be provided for laboratories to comment on operational problems and test validity. Ideally this should be in the last used column. It aids collation if the laboratory code and laboratory name are stored in the first two columns.

Close attention must be paid to how data is formatted, particularly date/time stamps and text fields (e.g. laboratory code, fluid name, instrument id, operator, validity codes, comments/problems, ...). Instructions must be given on the number of decimal places to which data is to be recorded. Clause 5.1.4 of ISO 5725-2 [1] recommends reporting to one more d.p. than specified in the test method. If the method does not specify the number of digits, then rounding shall not be coarser than half the repeatability SD. When precision may depend on performance level, different degrees of rounding may be needed for different levels. Working groups may have to use spreadsheets for data collation rather than the database if the data dictionary puts unacceptable constraints on digits recorded.

Example

The following template might be distributed to laboratory B. This indicates the order in which tests are to be conducted (but it is not essential to do this on the template as it could be time consuming for the DBA to prepare and distribute copies with different test orders to different laboratories).

Lab Code	Lab Name	Test ID	Instrument model / generation	Instrument Serial number / ID	Fuel (A or B or C or D)	Operator Name / ID	Test start date/time (dd/mm/yyyy hh:mm, hours 00 to 23)	Result (g/km) (1 d.p.)	Ambient temp (C) (1 d.p.)	Test valid (Y/N)	Comments regarding test operation or validity
B	Melchester test site				A						
B	Melchester test site				C						
B	Melchester test site				B						
B	Melchester test site				D						
B	Melchester test site				C						
B	Melchester test site				D						
B	Melchester test site				A						
B	Melchester test site				B						

The laboratory would then enter its results and complete the spreadsheet as follows:

Lab Code	Lab Name	Test ID	Instrument model / generation	Instrument Serial number / ID	Fuel (A or B or C or D)	Operator Name / ID	Test start date/time (dd/mm/yyyy hh:mm, hours 00 to 23)	Result	Ambient temp (C)	Test valid (Y/N)	Comments regarding test operation or validity
B	Melchester test site	LABO_001	Mark 3	M602	A	Bruce	25/12/2006 09:00	14.1	2.2	Y	
B	Melchester test site	LABO_002	Mark 3	M602	C	Bruce	25/12/2006 10:05	10.2	2.9	Y	
B	Melchester test site	LABO_003	Mark 3	M602	B	Bruce	25/12/2006 11:07	5.1	3.1	Y	
B	Melchester test site	LABO_004	Mark 3	M602	D	Bruce	25/12/2006 12:15	25.1	3.2	N	Invalid: engine knock
B	Melchester test site	LABO_005	Mark 3	M602	C	Bruce	25/12/2006 13:29	10.9	2.9	Y	
B	Melchester test site	LABO_006	Mark 3	M602	D	Bruce	25/12/2006 14:35	18.5	2.6	Y	
B	Melchester test site	LABO_007	Mark 3	M602	A	Bruce	25/12/2006 15:41	14	1.1	Y	
B	Melchester test site	LABO_008	Mark 3	M602	B	Bruce	25/12/2006 16:47	4.7	-0.5	Y	
B	Melchester test site	LABO_009	Mark 3	M602	D	Bruce	25/12/2006 17:58	18.9	3.2	Y	Repeat of LABO_004

Participating laboratories shall enter the complete set of test results generated during the round robin into the CEC-TMS or ATC-ERC database, or supply these to the WG Database Administrator via the spreadsheet provided, as appropriate, within an agreed time scale. All requested information must be provided and all problems both in individual tests and overall shall be recorded.

The WG DBA shall ensure that all results accepted as meeting the requirements of the round robin are collated into a single data set. The WG DBA shall forward this to the SDG Liaison Officer for analysis by a mutually agreed date in an agreed format. This is likely to be similar to the example above with results from the second laboratory following on below those from the first in the same column(s). The results from subsequent laboratories are then appended similarly.

The WG DBA shall ensure that test results that have been declared invalid are either clearly indicated in the data set or excluded completely. Tests should only be declared invalid if a deviation from the test procedure has been identified. Operational parameters outside tolerance limits or operational faults will normally invalidate the test and all its parameters. Missing parameters will not normally invalidate the test. **Data points shall not be removed by the WG DBA simply because they are out of line with the rest.** Invalid tests must be indicated by proper validity codes avoiding reliance on comments or formatting.

Once the due date has been passed, the SDG LO shall refuse to recycle precision analyses to incorporate overdue data.

Laboratories and stands are only identified by codes in the CEC-TMS and ATC-ERC databases. If laboratory names are required, then procedures should be put in place to identify laboratories, or to allow laboratories to identify their own results, commensurate with the data confidentiality rules pertaining within the Working Group at the time. Proforma 3 in the CEC Constitution Operating Guidelines provides a suitable Reference Test Acceptance and Identification Form for test results stored in main ATC-ERC engine test database.

8. Statistical Analysis

The SDG LO shall analyse the round robin data set that is either supplied by the WG DBA or extracted from the CEC-TMS or ATC-ERC database. The SDG LO shall not amend the data set without the agreement of the Working Group Chairman and/or DBA.

The SDG LO shall refuse to accept results arriving after the agreed deadline.

The prime objective is to determine the precision (repeatability and reproducibility) of each reported parameter in the test method and produce or update its precision statement.

If data is available from earlier round robins and/or test monitoring, the SDG LO should also look at trends in test precision and severity and report any significant changes to the Working Group. In addition, the SDG Liaison Officer may be asked to provide other analyses and advice that will be useful for improving the understanding of the test.

Precision analyses shall be conducted in accordance with SDG methodology and the procedures laid down in International Standard ISO 5725 part 2 [1]. There are some differences between SDG methodology and ISO 5725-2 and these are detailed in Appendix A.

The precision analysis should be conducted using appropriate statistical software from a recognised supplier. It is recommended that precision analyses be conducted using the START statistical program, supplied to CEC by Infineum. START is provided as a Microsoft Excel® add-in.

The SDG LO shall examine the data for potential outliers using raw data plots, laboratory comparison charts and statistical outlier tests, as described in ISO 5725-2 and Appendix A. Where possible, suspect data points should be queried with the originator via the WG DBA. When several unexplained abnormal test results occur within the same laboratory, the complete set of results from that laboratory may be queried. The SG LO and WG DBA shall then decide whether the suspect result(s) should be corrected, discarded or retained taking both engineering and statistical judgement into account, and record the reasons. Policies and decisions will also be needed on how the rejection of one reported parameter in a test affects the validity of others.

It is important to treat outliers in a consistent way in precision studies as retention/rejection decisions have a marked effect on precision estimates, particularly in smaller round robins. As a rule of thumb, data points or data cells that are significant at $P < 1\%$ in Grubbs' or Cochran's test (see ISO 5725-2) should be rejected as "outliers". Points or cells that are significant at $P < 5\%$ ("stragglers") should normally be retained unless there has been a material departure from the test procedure or normal test conditions.

For some CEC working groups, the test monitoring approach described in Procedure 2 is either premature, e.g. in the early stages of test development when the method is not sufficiently stable, or inappropriate, e.g. if samples have a limited life. Such working groups will need to consider whether laboratories that produce round robin results that are out of line with their peers are in statistical control. Procedure 2 Section 1.13 and Procedure 3 Appendix B give guidelines for assessing the acceptability of round robin results.

9. Reporting

The SDG LO shall prepare a written report documenting the results of the precision analysis and send this to the Working Group Chairman. Suitable formats include Microsoft Powerpoint®, Microsoft Word® and Adobe Acrobat® PDF. It is good practice to let the Working Group chairman see a draft of the report/analysis before issue. The results of the analysis shall be discussed at the next Working Group meeting and recorded in the minutes or its appendices. The Working Group Chairman may request the SDG LO to present the results in person.

Electronic data files containing the raw round robin data and the full numerical and graphical results of the precision analysis must be lodged in the Working Group area of the CEC website. An Excel file containing the raw data and the analyses and plots produced by the START program will usually suffice.

Unless agreed otherwise by the Working Group, reports shall code participating laboratories.

The report should assist the Working Group and CEC Management Board in determining the fitness for purpose of the test procedure and provide all the information that is needed for this assessment.

The report shall include the following statistics for each reported parameter for each sample tested:

Number of laboratories
Total number of tests
Mean
Repeatability SD *r SD*
Repeatability[4] *r*
Reproducibility SD *R SD*
Reproducibility *R*

as defined in Part 1 of ISO 5725 and Appendix A.

The report shall also compare the Repeatability *r* and Reproducibility *R* against their respective targets and calculate

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

Targets may be based on overall precision figures, particular samples or precision estimated from a functional relationship at some fixed performance level (see Procedure 3 for details).

[4] In CEC precision statements, the repeatability represents the likely difference between two test results on the same sample conducted within a short interval of time (same test conditions, same operator, same apparatus, same laboratory), but not back-to-back. Changes of sample are assumed to occur between the two repeat tests. See Appendix A for further details.

The report may also include overall repeatability and reproducibility figures calculated across samples as per ISO 5725-2 and Appendix A. When precision values vary with the mean, functional relationships may be established using the method described in section 7.5 of ISO 5725-2. Functional relationships should only be derived when sufficient samples have been tested to cover the range of applicability of the test and capture any nonlinearity. Precision estimates based on such functional relationships should not be extrapolated beyond the range of measured values for the samples tested.

Example

A typical precision summary produced by START is shown below. Some editing has been done to show the Reproducibility target and Q_R appropriately.

SG-L-040 2005 NOACK round robin (Procedure B)
Response: NOACK (1 Lab plus 1 Outlier Pair Excluded)

Reproducibility and Repeatability Estimates

Sample	Lab Count	No. Results	Mean	rSD	RSD	r	R
Oil A RL208/1	36	72	11.66	0.16	0.26	0.44	0.73
Oil B RL172/5	36	72	13.85	0.16	0.40	0.45	1.12
Oil C Noack CRM L-type W4520001	36	72	10.94	0.21	0.38	0.59	1.06
Oil D Noack CRM H-type W4520004	35	70	13.15	0.16	0.42	0.45	1.17
					Overall Repeatability =		0.49
					Overall Reproducibility =		1.03
Reproducibility Target (at 14%)	1.4		(based on RL172/5)			Q_R =	0.80

P.J. Zemroch RR2005 results CEC Noack (START) (proc B).xls 24/10/2005 SDG Statistical Analysis

Variance stabilising transformations provide an alternative to fitting functional relationships for deriving precision statements expressed as a function of the mean response. For example, if the repeatability standard deviation (say) is proportional to the mean performance level, then the observed values *y* can be converted to transformed values $z = \ln(y)$, and a precision analysis conducted on the values of *z.* If the repeatability of *z* turns out to be r_z, then the repeatability r_y on the raw y scale is $r_y = r_z \times y$, where *y* is the true performance level of the sample under test.

Detailed procedures for finding appropriate variance stabilising transformations may be found in International Standard ISO 4259 [2].

If the repeatability (or reproducibility) is reasonably constant for the different samples on the transformed *z* scale, then equations for the precision of an arbitrary sample may be derived from the overall repeatability or reproducibility figure, e.g. $r_y = r_z \times y$ as above. However such relationships should only be derived when sufficient samples have been tested to cover the range of applicability of the test. Such relationships should not be extrapolated beyond the range of average results for the samples tested in the round robin.

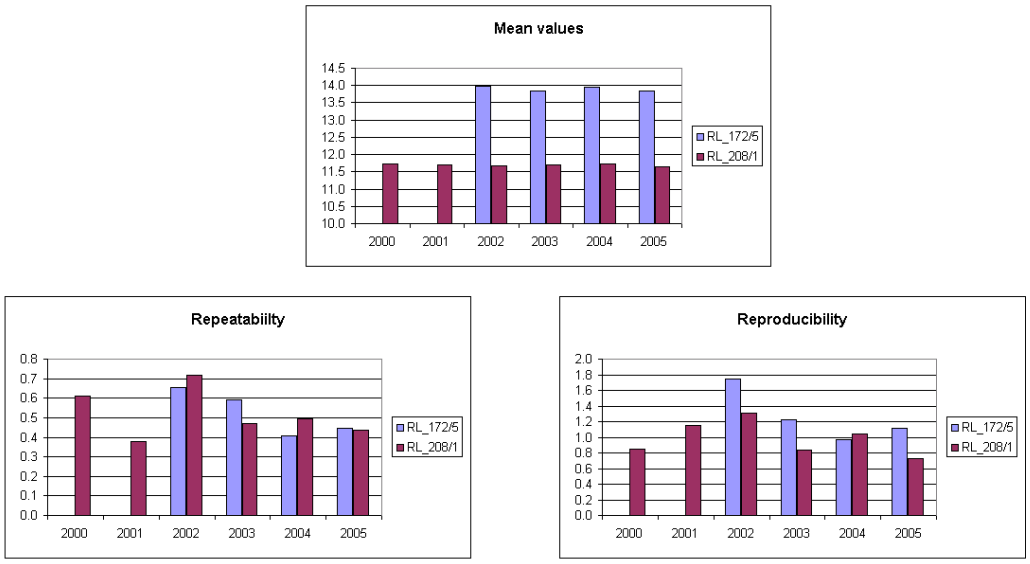
In some round robin exercises, there will be no simple relationship between the precision (repeatability and reproducibility) of the various samples and their mean performance level. For example, precision could depend on the chemical composition or the presence of additives. This should be noted both in the report to the working group and in the precision statement. Overall repeatability or reproducibility figures should not be quoted unless (a) a sufficient number of samples has been tested, representing the main types of test fluids falling within the scope of the method, (b) the results span the range of applicability of the method and (c) the precision values do not differ significantly from one sample to another.

When the round robin is one of a series of studies over a period of time, tables or graphs should be produced showing trends in mean values, repeatability and reproducibility.

Such tables and graphs may also include means and precision estimates based on test monitoring data. However care needs to be taken when comparing repeatability estimates derived from round robins with those derived from test monitoring. Test monitoring data are generally derived under “site precision” conditions (same sample, same laboratory, same instrument, extended periods of time, operators and other test conditions may vary) rather than “repeatability” conditions (same sample, same laboratory, same instrument, same operator, short intervals of time). Therefore repeatability estimates from round robins are not generally comparable with those from test monitoring. However rough estimates of shorter-term repeatability can be obtained from test monitoring data by looking at differences between successive values (see Procedure 2).

Example

The following plots show the results of SG-L-040 round robins conducted between 2000 and 2005.



Estimates of repeatability *r* and reproducibility *R* are subject to uncertainty in just the same way as estimates of means. Formulae for calculating confidence limits are given in Appendix B. The START program provides 95% confidence limits for the sample mean, and also for *r* and *R*, and these can be used to add error bars to plots of mean values and precision against time. These would help Working Groups visualise whether changes are significant or not.

Significant changes in mean values over time could be indicative of changes in test severity or changes in the condition of the test fluid batch. These must be brought to the attention of the Working Group. Similarly significant changes in precision must be reported.

Possible changes in means can be identified from tables or plots and confirmed using standard statistical techniques such as *t*-tests, analysis of variance or regression. Similarly changes in variability can be tested using *F*-tests, generalised linear modelling, or Bartlett's or Cochran's tests. Precision estimates in some round robins may be more accurate than in others, so proper account must be taken of round robin size which determines degrees of freedom. Statistical advice should be taken as estimates of means and standard deviations in successive years may not be truly independent if these emanate from broadly the same set of laboratories.

The SDG LO shall use the results of the round robin analysis to provide or review the precision statement for the test method, as defined in Procedure 3. The precision statement need not be based solely on precision estimates from

the current round robin; it may also incorporate estimates from other recent round robins or test monitoring. Where applicable the test monitoring targets and limits should also be reviewed (see Procedure 2).

The choice of data requires both statistical and engineering judgement. The incorporation of historical precision data is most likely to be appropriate when (a) no changes have been made to the test method, (b) the current round robin is small and (c) no significant changes have been seen in severity or precision over time. Test monitoring data can be used to improve estimates of reproducibility but should not normally be used to estimate repeatability as data is not collected under repeatability conditions.

10. 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.

Appendix A. Exceptions to International Standard ISO 5725

Repeatability and Site precision

ISO 5725 part 1 paragraph 3.13 [1] defines “repeatability conditions” as “Conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time”.

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. 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.

“Site precision conditions” are defined in standard ASTM D6299 [3] as “Conditions under which test results are obtained by one or more operators in a single site location practicing the same test method on a single measurement system using test specimens taken at random from the same sample of material, over an extended period of time, spanning at least a 15-day interval.” Test monitoring data are normally collected under site precision conditions.

“Site precision” is the value equal to or below which the absolute difference between two single test results obtained under site precision conditions may be expected to lie with a probability of 95%.

Statistical analysis of round robin data

CEC/SDG precision analyses are conducted in accordance with Part 2 of International Standard ISO 5725 with the exceptions listed below.

7.2.2 Redundant data – In situations where laboratories carry out and report more tests than requested on a particular sample or samples, then normally all valid results will be included in the analysis. However this may impact on the validity of Cochran’s and Grubbs’ tests and so statistical advice needs to be taken.

7.3.3.6 Cochran’s test – This paragraph requires the deletion of entire laboratory × sample cells if the standard deviation is classed as an outlier in Cochran’s test. There may be situations where it is clear which result within a cell is the outlier. In such circumstances that result alone may be discarded.

7.2.4/7.2.5/7.3.1 Outliers / Outlying laboratories / Graphical consistency techniques

The first step in the search for outliers and outlying laboratories is to plot the raw data, for example, as shown in Figure 1.

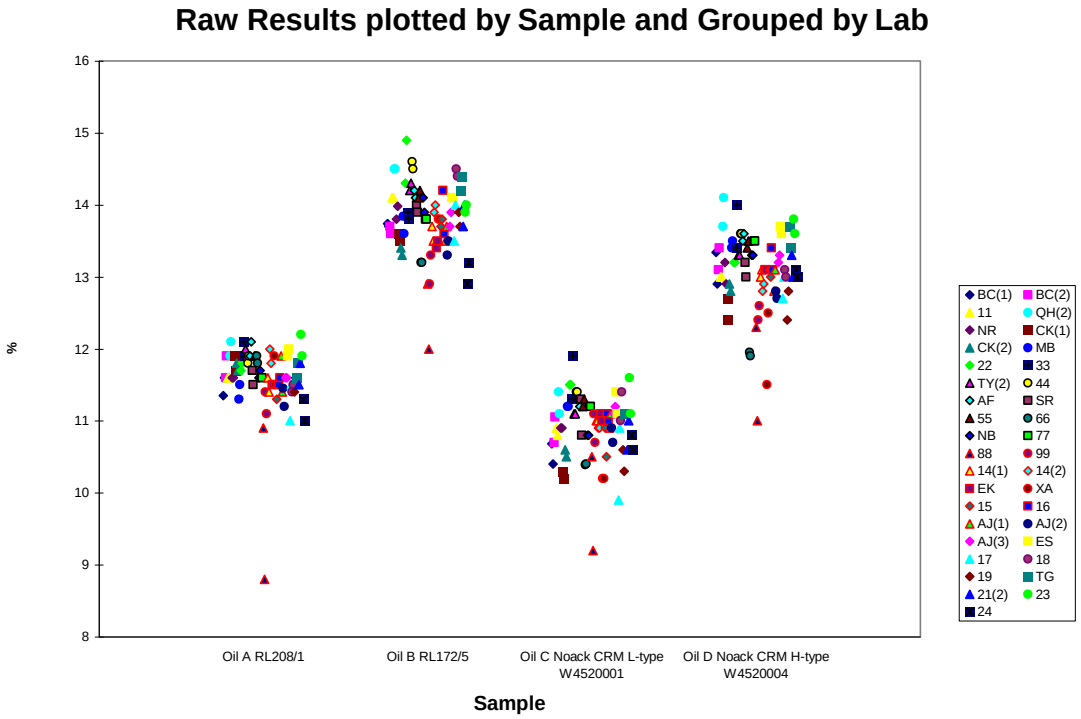


Figure 1. Typical raw data plot produced by START

Obviously erroneous data points should be investigated and corrected or discarded as per paragraph 7.2.6. When several unexplained abnormal test results occur form the same laboratory, for example laboratory 88 in Figure 1, then that laboratory may be considered an outlier and all its results rejected.

Mandel's *h* and *k* statistics are not calculated in the START program. Instead laboratories may be compared by means of "*Laboratory comparison charts*".

Laboratories can differ from their peers in terms of

(a) severity (do they measure systematically high or low relative to their counterparts?),
(b) the way in which they rank different fluids or
(c) repeatability.

Outlying laboratories are identified using raw data plots and laboratory comparison charts. Appendix C describes how laboratory comparison charts may be constructed and interpreted. Figure 2 shows the laboratory comparison chart for the data Figure 1. This highlights the serious severity problems at laboratory 88.

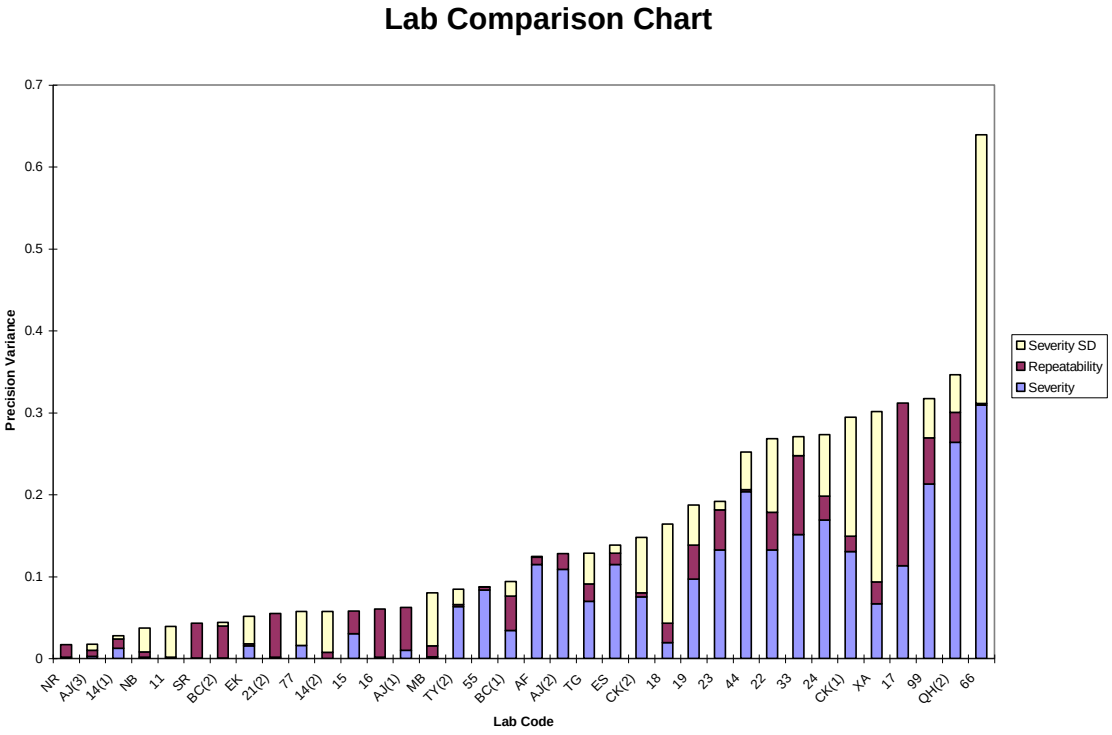


Figure 2. Example of Laboratory Comparison Chart

7.4.4 Calculation of the general mean $\hat{m}$

In CEC calculations, each laboratory is given equal weight when calculating the mean performance level for any particular sample, irrespective of the number of tests conducted. Thus

$$\hat{m}_j = \sum_{i=1}^{p} \left(\sum_{k=1}^{n_{ij}} y_{ijk} / n_{ij} \right) / p$$

7.4.5.4 Negative variance components

In some CEC round robin analyses, the estimate of the between-laboratory variance s_{LJ}^2, obtained by the methods in section 7.4.5.3 of ISO 5725, turns out to be negative owing to random effects. This problem occurs where inter-laboratory effects are modest and is particularly prevalent in smaller round robins where the number of degrees of freedom for repeatability is small (e.g. Table 3 in Section 5).

ISO 5725 recommends assuming s_{LJ}^2 to be zero in such circumstances. However this can lead to very unreliable estimates of precision with both the repeatability and reproducibility being estimated from the repeatability variance s_{rJ}^2 which may be subject to high levels of uncertainty. Variations between test results at different laboratories are, in effect, ignored.

Statistical advice needs to be taken in such situations. One approach might be to simply calculate the standard deviation of all test results on a particular sample, regardless of where the test results were conducted. This would lead to a single measure of precision refelecting the closeness of agreement between pairs of results wherever they may be taken. Another approach might be to use Bayesian methods as described in Gilmour and Goos[5]. These are implemented in the WinBUGS freeware package (see http://www.mrc-bsu.cam.ac.uk/bugs/ and Lunn, Thomas, Best and Spiegelhalter[6]).

7.5 Variance stabilising transformations

Section 7.5 of standard ISO 5725 part 2 describes how functional relationships can be established between the repeatability and/or reproducibility of a test method and the mean performance level using weighted regression techniques. However the number of functional forms considered is limited.

A wider range of functional relationships can be derived by means of variance stabilising transformations. Detailed procedures for finding appropriate functional forms may be found in International Standard ISO 4259 [2].

7.6.14 Calculation of the overall repeatability r and reproducibility R

CEC calculates the overall repeatability and reproducibility across samples by averaging the corresponding variances rather than the standard deviations (see Appendix B for details).

[5] Gilmour, S. G. and Goos, P. (2009) Analysis of data from nonorthogonal multi-stratum designs in industrial experiments. Journal of the Royal Statistical Society, Series C, 58, 467-484.

[6] Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325-337.

Appendix B. Approximate Degrees of Freedom and Confidence Intervals for Sample Means, Repeatability and Reproducibility

International Standard ISO 5725-2 [1] and Appendix A give methods of estimating the sample-by-sample mean $\hat{m}_j$, repeatability variance s_{rj}^2 and reproducibility variance s_{Rj}^2 for a test method from round robin data. However, the standard gives no means of estimating the uncertainty in these values. The repeatability r_j and reproducibility R_j for each sample j = 1, 2, ..., p are calculated from the corresponding variances using the formulae

$$r = 2.8 \times s_r \qquad\qquad R = 2.8 \times s_R ,$$

but again there is no written procedure for estimating uncertainty.

Let us assume that a round robin has been conducted and that laboratory i (i = 1, 2, ..., p) has collected n_{ij} independent measurements y_{ijk} on sample j. Then, if the values of y_{ijk} are normally distributed, the measured repeatability variance s_{rj}^2, given by equation (20) in ISO 5725-2, will be distributed as $\sigma_{rj}^2 \chi_{\nu_{rj}}^2 / \nu_{rj}$ where σ_{rj}^2 is the true repeatability variance and the number of degrees of freedom ν_{rj} is given by

$$\nu_{rj} = \sum_{i=1}^{p_j} (n_{ij} - 1),$$

where the summation is over the p_j laboratories reporting at least one valid measurement on sample j.

A 95% confidence interval for the true repeatability variance σ_{rj}^2 is then

$$\frac{s_{rj}^2 \nu_{rj}}{\chi^2_{\nu_{rj};0.975}} < \sigma_{rj}^2 < \frac{s_{rj}^2 \nu_{rj}}{\chi^2_{\nu_{rj};0.025}}$$

where $\chi^2_{\nu_{rj};0.975}$ and $\chi^2_{\nu_{rj};0.025}$ are the upper and lower 2.5% points (one-sided) of the χ^2-distribution with ν_{rj} d.f. It follows that a 95% confidence interval for the true repeatability r is

$$2.8 s_{rj} \sqrt{\frac{\nu_{rj}}{\chi^2_{\nu_{rj};0.975}}} < r < 2.8 s_{rj} \sqrt{\frac{\nu_{rj}}{\chi^2_{\nu_{rj};0.025}}}$$

In section 7.4.5 of ISO 5725-2, the reproducibility variance s_{Rj}^2 for sample *j* is effectively calculated from the between-laboratory mean square s_{dj}^2 and the repeatability variance s_{rj}^2 using the equation

$$s_{Rj}^2 = \frac{s_{dj}^2}{\bar{\bar{n}}_j} + s_{rj}^2\left(\frac{\bar{\bar{n}}_j - 1}{\bar{\bar{n}}_j}\right)$$

where s_{dj}^2 has $p_j - 1$ degrees of freedom. It can be shown that the measured reproducibility variance s_{Rj}^2 above is approximately distributed as $\sigma_{Rj}^2\chi_{\nu_{Rj}}^2 / \nu_{Rj}$ where σ_{Rj}^2 is the true reproducibility variance and

$$\nu_{Rj} = \frac{(s_{Rj}^2)^2}{\frac{(s_{dj}^2 / \bar{\bar{n}}_j)^2}{p_j - 1} + \frac{((\bar{\bar{n}}_j - 1)s_{rj}^2 / \bar{\bar{n}}_j)^2}{\nu_{rj}}}.$$

An approximate 95% confidence interval for the true reproducibility variance σ_{Rj}^2 is then

$$\frac{s_{Rj}^2\nu_{Rj}}{\chi_{\nu_{Rj};0.975}^2} < \sigma_{Rj}^2 < \frac{s_{Rj}^2\nu_{Rj}}{\chi_{\nu_{Rj};0.025}^2}$$

and an approximate 95% confidence interval for the true reproducibility *R* is

$$2.8s_{Rj}\sqrt{\frac{\nu_{Rj}}{\chi_{\nu_{Rj};0.975}^2}} < R < 2.8s_{Rj}\sqrt{\frac{\nu_{Rj}}{\chi_{\nu_{Rj};0.025}^2}}.$$

(Note: the above approximations are likely to be less accurate for small data sets).

The variance of the sample mean $\hat{m}_j$, calculated as per Appendix A giving each laboratory equal weight, is

$$\mathrm{var}(\hat{m}_j) = \frac{\sigma_{Lj}^2}{p_j} + \sigma_{rj}^2\frac{\sum_{i=1}^{p_j}(1/n_i)}{p_j^2}.$$

where σ_{Lj}^2 is the true between laboratory variance. Var($\hat{m}_j$)is estimated as

$$\mathrm{var}(\hat{m}_j) = \frac{{s_{dj}}^2}{p_j n_j} + \frac{{s_{rj}}^2}{p_j}\left(\frac{\sum_{i=1}^{p_j}(1/n_i)}{p_j} - \frac{1}{n_j} \right)$$

The approximate degrees of freedom ν_j for this estimate are

$$\nu_j = \frac{\mathrm{var}(\hat{m}_j)^2}{\frac{(s_{dj}^2 / n_j)^2}{p_j - 1} + \frac{1}{\nu_{rj}}\left({s_{rj}}^2 \left(\frac{\sum_{i=1}^{p_j}(1/n_i)}{p_j} - \frac{1}{n_j} \right) \right)^2}$$

A 95% confidence interval for the true mean value can then be calculated as

$$\hat{m}_j \pm t_{\nu_j;\,0.05} SE(\hat{m}_j)$$

where $SE(\hat{m}_j) = \sqrt{\mathrm{var}(\hat{m}_j)}$ is the standard error of the sample mean $\hat{m}_j$, and $t_{\nu_j;0.05}$ is the upper 5% point (two sided) of the t-distribution with ν_j d.f.

Overall repeatability and reproducibility

Now let us suppose that the repeatability r_j and reproducibility R_j do not vary in any consistent way with the sample mean m_j. ISO 5725-2 then allows a pooled estimate of the overall repeatability standard deviation s_r to be formed by simply averaging the repeatability S.D.s $s_{r1}, s_{r2}, \ldots, s_{rq}$ for each of the *q* samples. This procedure departs from the standard at this point and averages the repeatability variances s_{rj}^2 instead, i.e.

$$s_r^2 = \frac{s_{r1}^2 + s_{r2}^2 + \ldots + s_{rq}^2}{q}.$$

Both the ISO 5725 estimate and the estimate s_r^2 above disregard any differences in the degrees of freedom available to estimate the *q* values s_{rj}^2 due to imbalance in the design, missing values or outlier rejection.

It can be shown that the overall repeatability variance s_r^2 above is approximately distributed as $\sigma_r^2 \chi_{\nu_r}^2 / \nu_r$ where σ_r^2 is the true overall repeatability variance and

$$\nu_r = \frac{(s_r^2)^2}{\frac{(s_{r1}^2 / q)^2}{\nu_{r1}} + \ldots + \frac{(s_{rq}^2 / q)^2}{\nu_{rq}}}.$$

This procedure departs similarly from ISO 5725 when calculating the overall reproducibility variance s_R^2 by recommending the average of the sample-by-sample reproducibility variances s_{Rj}^2, i.e.

$$s_R^2 = \frac{s_{R1}^2 + s_{R2}^2 + \ldots + s_{Rq}^2}{q}.$$

Splitting each term s_{Rj}^2 into contributions from the between-laboratory mean square s_{dj}^2 and repeatability variance s_{rj}^2, we obtain

$$s_R^2 = \frac{s_{d1}^2}{\bar{=}qn_1} + \ldots + \frac{s_{dq}^2}{\bar{=}qn_q} + \frac{s_{r1}^2}{q}\left(\frac{\bar{=}n_1 - 1}{\bar{=}n_1}\right) + \ldots + \frac{s_{rq}^2}{q}\left(\frac{\bar{=}n_q - 1}{\bar{=}n_q}\right).$$

The between-laboratory mean squares s_{dj}^2 will not be mutually independent unless each sample *j* is tested at a different set of p_j laboratories. As no sensible or practicable round robin would be conducted in this way, we CANNOT use the obvious equation

$$\nu_R = \frac{(s_R^2)^2}{\frac{(s_{R1}^2 / q)^2}{\nu_{R1}} + \ldots + \frac{(s_{Rq}^2 / q)^2}{\nu_{Rq}}}$$

to estimate the degrees of freedom ν_R for the overall reproducibility *R* unless laboratory-to-laboratory variation is totally absent.

If all *q* samples are tested at the same *p* laboratories, and if laboratory-to-laboratory variation is large, then the between-laboratory mean squares s_{dj}^2 will each be dominated by the systematic variations between these *p* laboratories. In such circumstances, it is reasonable to treat the sum $s_{d1}^2 / q\bar{=}n_1 + \ldots + s_{dq}^2 / q\bar{=}n_q$ as a single entity with approximately $p-1$ degrees of freedom. This combined term and the repeatability variances s_{rj}^2 are mutually independent, so the overall reproducibility variance s_R^2 above is approximately distributed as $\sigma_R^2 \chi_{\nu_R}^2 / \nu_R$ where σ_R^2 is the true overall reproducibility variance and the overall reproducibility degrees of freedom ν_R are estimated by

$$\nu_R = \frac{(s_R^2)^2}{\frac{\left(\frac{s_{d1}^2}{\overline{\overline{qn_1}}} + \ldots + \frac{s_{dq}^2}{\overline{\overline{qn_q}}}\right)^2}{p-1} + \frac{\left[\frac{s_{r1}^2}{q}\left(\frac{\overline{\overline{n_1}}-1}{\overline{\overline{n_1}}}\right)\right]^2}{\nu_{r1}} + \ldots + \frac{\left[\frac{s_{rq}^2}{q}\left(\frac{\overline{\overline{n_q}}-1}{\overline{\overline{n_q}}}\right)\right]^2}{\nu_{rq}}}.$$

Approximate 95% confidence intervals for the true overall repeatability variance σ_r^2, reproducibility variance σ_R^2, repeatability *r* and reproducibility *R* can be now derived as in the single sample case.

Appendix C. Laboratory comparison charts

The relative performances of the various laboratories participating in CEC Round Robins can be visualised by means of "Laboratory comparison charts". These rank laboratories according to three criteria:

Overall Severity (S1)

This is a measure of whether a laboratory produces test results that are consistently high or consistently low relative to its peers. There could be many reasons for this, e.g. differences in test engines, different batches of key parts, differences in installation, systematic differences in measurement, etc.

Repeatability (S2)

This is a measure of a laboratory's ability to obtain similar results when retesting the same sample. Poor repeatability could come from inconsistencies in setting up and running the test, measurement errors, changes in ambient conditions, etc.

Severity SD (S3)

This is a measure of a laboratory's ability to rank samples in the same order as its peers. The Severity SD is independent of the Overall Severity and Repeatability. For example, a laboratory might not be considered particularly severe or mild overall, and it might obtain repeatable results. However it could still rank samples in a different order to other laboratories. A high Severity SD could represent real differences in the way a laboratory treats different types of samples, or it might just be a consequence of one or more extreme results.

These components are then combined to give an overall precision value for each laboratory, the higher the value the worse the precision. This is referred to as the Total SD value and is used to rank laboratories.

Figure 3 below shows a laboratory comparison chart produced by the START program, based on the raw data.

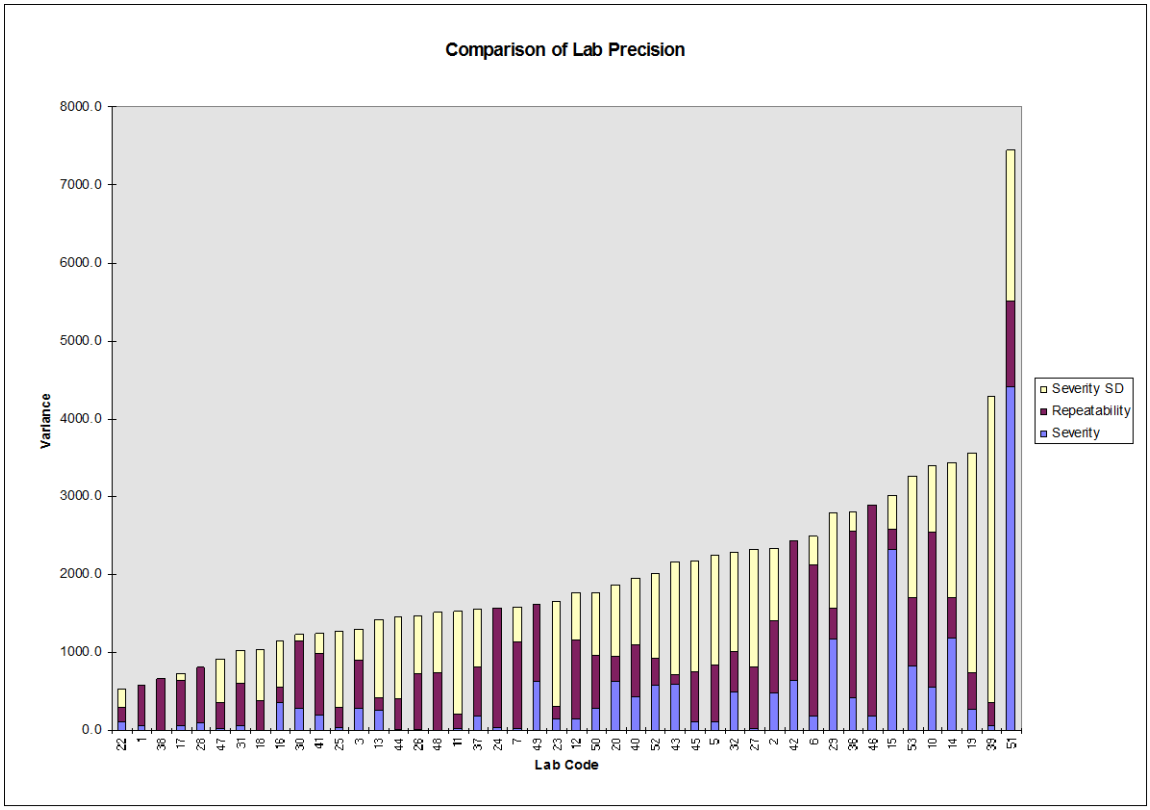


Figure 3. Graphical Comparison of Laboratories in the 1996 PF-006 lubricity round robin.

In this example, laboratory 51 has the worst precision overall, and the prime cause is its overall severity. The accompanying table (see extract below) indicates that laboratory 51 gives systematically lower values than its peers which, in this test, means it is “mild” rather than “severe”.

Lab	Severity	r SD	Severity SD	Total SD
22	-10	14	15	23
38	-1	26	0	26
17	-8	24	9	27
…	…	…	…	…
44	-6	58	21	62
39	-8	17	63	65
51	-75	57	28	98

Laboratories 19 and 39 have poor overall precision due to a high Severity SD, even though they have good repeatability and are neither mild nor severe overall. Indeed overall severity is not a major concern at most laboratories. The main issues appear to be poor repeatability and inconsistent ranking of test fluids.

Calculations

The calculations below may either be performed on the raw or standardised data. The standardised data are obtained by dividing each raw result by the estimated reproducibility Standard Deviation for the sample.

It is recommended to standardise when the reproducibility standard deviation varies across samples. Since this is the case more often than not, standardising is the default method implemented within START. An alternative approach is to apply a variance stabilizing transformation to the data (ISO 4259 [2] provides some examples).

Overall Severity (S1)

The Overall Severity for laboratory *i* is calculated as

$$S1_i = \sum_{j=1}^{q} \frac{d_{ij}}{q}$$

where

$$d_{ij} = \bar{y}_{ij} - M_j ,$$

$\bar{y}_{ij}$ is the mean value for sample *j* at laboratory *i* and M_j the median of the cell means $\bar{y}_{ij}$ across the *p* laboratories.

Note: The median M_j is used in preference to the mean of the $\bar{y}_{ij}$'s in the above formula to provide robustness against extreme results which are not removed as outliers.

Repeatability SD (S2)

This repeatability SD for laboratory *i* is the pooled within sample standard deviation, giving equal weight to all samples tested more than once, regardless of the number of tests conducted:

$$S2_i = \sqrt{\sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} \frac{(y_{ijk} - \bar{y}_{ij})^2}{n_{ij} - 1} / q}$$

where n_{ij} is the number of tests conducted on sample *j* at laboratory *i* (cf Appendix B).

Severity SD (S3)

The severity SD is calculated as

$$S3_i = \sqrt{\text{var}(d_{1j,}d_{2j},\ldots d_{pj}) - \frac{S2_i^2}{q/\sum_{j=1}^{q}\frac{1}{n_{ij}}}}$$

If the term underneath the square root sign is negative, then $S3_i$ is taken as zero. If a sample is not tested at a particular laboratory, then it should be excluded from the calculation of $S3_i$.

Total SD

The Total SD combines the three sources of imprecision above and is calculated as

$$\text{Total SD} = \sqrt{S1_i^2 + S2_i^2 + S3_i^2}$$

The squared values S_1^2, S_2^2 and S_3^2 can then be plotted as a stacked bar chart, as in Figure 3. The total height of the bar will be the Total SD value squared.

Treatment of Outliers

The position of a laboratory in a comparison chart can depend on how potential outliers from that laboratory are treated. If the potential outliers are retained, then the laboratory will appear to do badly. If they are removed, then a laboratory with several outliers may appear to perform very well. Two options can be considered.

(1) Do the analysis with and without outliers
(2) Report the analysis with outliers removed, but note the numbers of outliers removed from each laboratory when reporting and interpreting the results.