Procedure 2

Test Monitoring Systems, Analysis of Test Monitoring Data and On-Line Data Depositories

1 Test Monitoring

1.1 What is Test Monitoring?

Test monitoring is the regular testing of samples with accepted reference values (i.e. check standards) by individual test laboratories in order to monitor their performance and allow remedial action to be taken when problems are detected. Test monitoring also allows the detection of industry-wide trends by collating the individual laboratories' data.

In order to maintain a high level of quality, CEC makes certain requirements relating both to the process of generating a test result, as described in CEC Guideline 18, and to the test result itself. Test monitoring systems are the tool that CEC uses to assess the uniformity of testing across laboratories, and the data will normally be stored centrally in either the CEC-TMS or ATC-ERC databases.

The results from test monitoring are plotted on control charts, an example of which is shown in Figure 1.

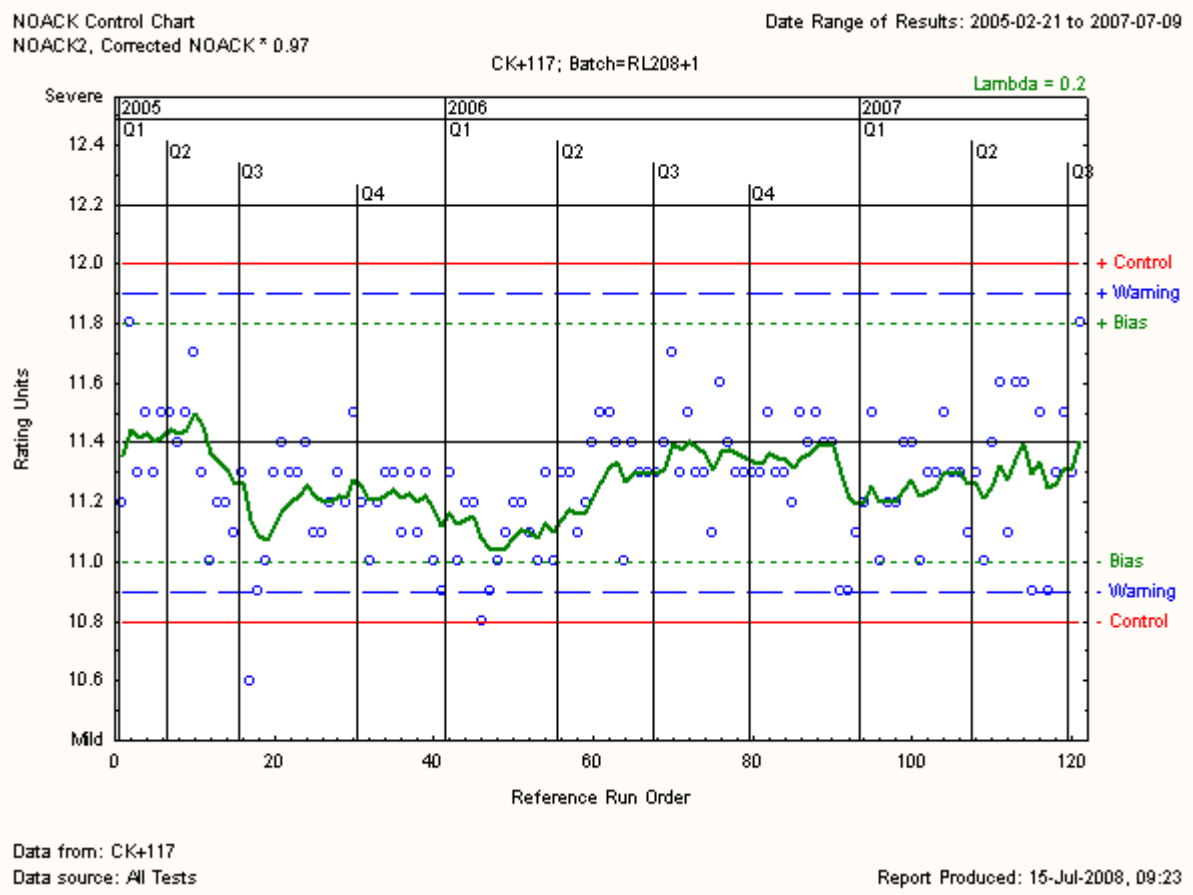

Figure 1. Example of a Control Chart generated by the CEC-TMS Test Monitoring System

In Figure 1 each dot represents the result of a single test. In this chart all results are from a single instrument, on a single oil sample, although this need not be the case. The sample used has had its performance evaluated in a round robin. From this a target value, shown by the horizontal black line, has been determined as well as the control, warning, and bias limits, shown in red, blue and green respectively.

In Figure 1 we see that one of the results is below the lower red control limit. This result is unacceptable, and subsequently a satisfactory result would need to be obtained before using the instrument to evaluate the performance of "candidate" oils of unknown performance. We say that an instrument is "out of control" if its most recent results are unacceptable, or "in control" if they meet the defined requirements.

The solid green line in Figure 1 smoothes the raw data points and is referred to as the trend line. The bias limits, plotted as horizontal green dashed lines, show permissible values for the trend line. Since the majority of the results for this instrument are below the target, it indicates that these results are somewhat lower than the industry average. However, since the trend line stays within the bias limits, this can be considered to be within the normal variability of the test.

The blue warning limits on a control chart can be used to warn a laboratory to carry out additional checks (see section 1.9).

1.2 Setting up a Test Monitoring System

Test monitoring will typically start when a method has satisfactorily completed both the test development stage and at least one full round robin amongst all participating laboratories. Test monitoring should only be used when a method is considered stable and no major changes in severity or precision are being observed. The results from the most recent round robin(s) will be used to determine the performance of the allocated reference samples, and to set the control, warning and bias limits for each. The requirements for laboratories to generate and submit reference data should then be written into section 11 of the WG's (working group's) test procedure.

The key elements involved in setting up a control chart based test monitoring system are described in sections 1.3 to 1.10 below. These include:

- Selection of parameters to be monitored
- Selection of reference samples to be tested
- Frequency of reference testing
- Calculating the mean and standard deviation for each sample
- Setting the target and the control, warning and bias limits
- Deciding upon any optional run rules
- The actions a laboratory needs to take when a limit is broken

1.3 Parameters to be monitored

The monitored parameters should normally be the parameters which the test method requires to be reported. These will usually include the primary parameters reflecting the prime function and purpose of the test as defined by the Management Board (see section 1 of Procedure 3), e.g. engine deposits, end of test viscosity, or wear on some mechanical component, and perhaps some "secondary" parameters serving a different purpose; e.g. safety or "no-harm" measures such as cylinder liner wear or bore polish.

In some circumstances, it may be desirable to plot other parameters, e.g. those that determine test validity, such as operational temperatures. These parameters should be used only for information and not for determining whether an out of control event has occurred.

In some tests, it may be decided that some of the reportable parameters are unsuitable for test monitoring and omitted. These might, for example, include parameters measured on a non-numeric or very coarse measurement scale.

Under some circumstances it may be considered desirable to monitor a derived quantity rather than the raw parameter measurement. However, this should be weighed against the benefit of simplicity. For example, a transformation, e.g. log(result), might be used when the distribution of repeat results is very asymmetric, whereas a difference between reference sample results might be used if it is desired to monitor discrimination. The SDG officer should provide advice on this.

1.4 Reference Samples to be tested

Reference samples should be selected to cover the critical levels of performance for each parameter in the test. Exceptionally, for a test with a single pass/fail level, this may be accomplished by using single products at the pass/fail level for each parameter. Normally, two or more products encompassing a range of performance on each parameter would be expected.

Criteria for selecting, sourcing, distributing and storing reference samples need to be established. In particular, reference samples should be stable and available in large batches. In order to prolong the usable life of a batch, special storage conditions may also need to be defined.

Test monitoring relies on laboratories being able to produce long sequences of repeat results on batches of known performance. If batches cannot be produced in sufficient quantities to allow this to happen then the group may have to rely on regular round robins to monitor its performance. For example, certain fuels tests require large quantities of fuel, meaning that one batch can only support a limited number of reference tests. Additionally, particular stability issues may arise with reference fuels containing biological components.

1.5 Frequency of Reference Testing

The typical frequency is one reference test followed by a maximum of nine candidate tests. If more than one reference product is being tested then the order of testing should be specified. For example, with two samples, a typical requirement is that testing should alternate between samples. Normally a maximum time between a candidate and the previous reference test, and between two consecutive reference tests, will also be stipulated.

1.6 Calculating the Mean and Standard Deviation

In order to determine realistic control limits, it is necessary to first calculate the mean and standard deviation of the parameter, for each reference sample. Initially the calculations will be based on round robin data; subsequently, data generated by the test monitoring system may be used.

For round robin data, the calculations are described in Procedure 1, while sections 2.4 and 2.5 of this procedure detail the process for test monitoring data. Care must be taken to remove suspect results. It is acceptable in this case for more results to be discarded as outliers than would be permitted for precision analyses. Only the reproducibility standard deviation (referred to as the SD in this procedure) and the mean of these cleaned data are used for control charts.

If it is necessary to introduce a new batch of a reference fluid, then a mini round robin exercise should normally be carried out in which each laboratory tests the new batch once. These results are used to determine the mean and reproducibility SD for the new batch, which can then be used to update the target and control chart limits. The precision statement may also be updated. If the repeatability SD is required then each laboratory must test the new batch twice.

If a mini round robin is not possible, then some laboratories should carry out back-to-back testing of the old and new batches in order to determine if there is any difference in performance. If even this is not possible, then the new batches should be validated against the existing control limits. However, the latter methods increase the risk that any out of control signals are due to batch changes rather than being laboratory related.

If it is necessary to introduce an entirely new reference fluid then a round robin exercise should normally be carried out in which each laboratory tests the new fluid at least twice. If this is not possible then each laboratory should test the new fluid at least once.

There is a danger that a change of reference fluid can lock in a temporary shift in test severity, seen during the time of the round robin. For this reason, if the new reference fluid is replacing another reference fluid, the old fluid should be tested back to back with the new fluid. This is mandatory where the specification is relative to the reference, except in the case that the old fluid is no longer available.

1.7 Calculation of the trend line for a single sample and single instrument or laboratory

Figure 2 shows an example of a control chart that might be manually constructed by a laboratory. The horizontal lines are the target and the various limits defined in the test procedure. Individual results are shown as disconnected points.

The trend line, or Exponentially Weighted Moving Average (EWMA) line, is calculated as follows.

The 0^{th} value of the EWMA is set to the target. If the i^{th} individual result on the chart is X_i then the i^{th} EWMA value ($EWMA_i$) is calculated using a smoothing parameter λ as:

$$EWMA_i = (1-\lambda)\, EWMA_{i-1} + \lambda X_i$$

When looking at data from a single laboratory or instrument, the smoothing parameter λ is typically set to 0.2. However, where references are produced infrequently (<10 over the life of a test installation/engine), $\lambda = 0.3$ is recommended.

If the EWMA goes outside the bias limits at the i^{th} result then, after an internal check, a further reference should be run. The $(i+1)^{th}$ value for the trend line should then be calculated as if restarting from the bias limit that was exceeded using the formula:

$$EWMA_{i+1} = (1-\lambda)\, \text{Bias Limit} + \lambda X_{i+1}$$

where X_{i+1} is the result of the retest.

If the trend line remains outside the bias limit, then a further check must be made.

If the check shows that a reference run was invalid, then that result should be removed from the calculation of the EWMA.

The EWMA points $EWMA_i$ should be consecutively joined by straight lines to give the trend line. When a control chart is plotted showing only the most recent data, then the EWMA should not start at the target but at the EWMA calculated from the preceding data.

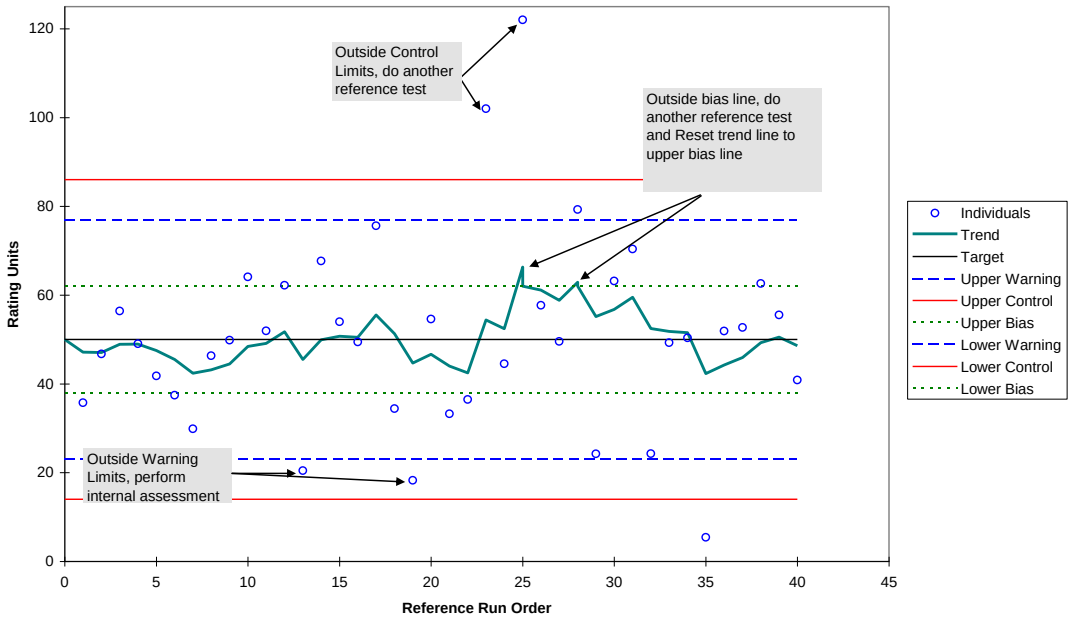

Figure 2. Example of a Control Chart created using a Spreadsheet

The EWMA should be restarted from the target when a major hardware change takes place, for example, when a new test engine is installed on an existing stand. When a new batch of reference fluid is introduced, the working group will need to decide whether the EWMA should be restarted from the new target or simply continue as if no change had occurred (but with the new limits). This decision will require both statistical and engineering judgement. The procedure should describe any circumstances which might require the EWMA to be reset.

1.8 Setting the Target and the Control, Warning and Bias Limits

Setting the target and limits is the responsibility of the CEC Working Group (WG).

Target

The target is normally set to the mean, determined as per section 2.4, but may be rounded to reflect the number of significant digits recorded when the test result is reported.

Control Limits

It is recommended that the control limits be set at Target ± K SD, where K is in the range 1.6 to 2.0 (typically 1.8) and SD is the reproducibility SD. Higher values of K give less sensitive control limits, which may fail to detect problems. However, some reasons for choosing a higher value of K may be as follows:

- There is uncertainty in the estimates of the mean and SD, possibly due to these being based on a small data set
- More than one parameter is being monitored

- The parameter has excellent precision and excursions a little greater than 2 standard deviations are of little practical significance.

Lower values of K give tighter limits, and will cause more re-testing. However a smaller value may be chosen if the precision for the parameter is poor (the SD is large) and the group is working hard to improve it.

Warning Limits
When warning limits are set (see section 1.9 below), the recommended choice is Target ± W SD where W = 0.75 K.

Bias Limits
The trend line smoothes the individual points and should therefore show less variability than the original data. The bias limits determine how far the trend line can deviate from the mean before an action is required. The bias limits are set at Target ± B SD. It is recommended that B is a minimum of 1.0 and no larger than W (0.75K). A low value of B is likely to generate actions at any laboratory that systematically deviates from the target. Higher values of B allow larger systematic differences between laboratories without giving rise to actions and warnings. Higher values of B are necessary for methods where the reproducibility is substantially larger than the repeatability.

The control, warning and bias limits will normally be rounded to the same number of significant digits as the target. Measured values on the limits are considered to be inside the limit. Minor adjustments may be made to the target and limits to ensure symmetry, but care needs to be taken to ensure that the adjusted limits are neither too stringent nor too slack.

Particular care should be taken when the parameter can only take a small number of discrete values, for example, ratings having an integer scale (0, 1, …, 10). Test monitoring based on control charts is not appropriate for parameters which have a binary outcome, e.g. pass/fail, as it is difficult to set meaningful control limits.

If any of the targets or control, warning or bias limits are reset, then the values in force on the day that a particular test started should be used for that test.

1.9 Optional Run Rules

In addition to the standard rules deeming laboratories out of control if they exceed the control or bias limits, it may be desired to have additional “run rules” based on the warning limits.

Two common run rules which would render a laboratory/instrument out of control are:

- 2 out of 3 consecutive results outside one of the warning limits i.e. both high or both low.
- 2 out of 3 consecutive results outside either of the warning limits

A variation is to replace “2 out of 3 consecutive results” by “2 consecutive results” in either of the rules above.

1.10 Detecting and Responding to Violations of the Limits

Laboratories are responsible for determining whether their results meet the criteria laid down by the working group. In the CEC-TMS database, the on-line graphs will show whether results meet the control and/or warning limits. However, if the plotted trend line is very close to one of the bias limits, then it may be difficult to tell if a control rule has been broken. In cases where doubt exists, the exact numerical data should be downloaded and examined to determine if the trend line crossed the bias limit and should be reset. Note that the EWMA values should never be rounded.

An “action” is triggered when a control limit is exceeded, the trend line goes outside the bias limits or a run rule is violated. After an action, laboratories should follow the guidelines given in section 11 of the test procedure. As a minimum, this will mean that the laboratory has to carry out some checks and obtain an acceptable result when repeating the test.

A “warning” is triggered when a single result is obtained which is outside the warning limits, but does not trigger an action. After a warning, laboratories should check their test set up following any advice given in section 11 of the test procedure. They may then continue to test candidates.

The Working Group should develop policies, on a case by case basis, on how “actions” and “warnings” for one reported parameter in a test affect the control status of others.

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.

As a result of test monitoring it is possible that on an industry wide basis, one or more parameters, but not all, are declared “Out of Control” (see section 1). In this situation, the test may not be used to demonstrate conformance vis-à-vis the out-of control parameter in any specification until that test parameter is declared back “In Control”. However all parameters must continue to be recorded in the test monitoring database.

1.11 Creating Control Charts from multiple samples/batches

If there is more than one reference product, or if there are several batches from the same reference, it is recommended that additional charts pooling data across all samples or all batches are considered. Often a laboratory can be consistently high or low across all samples and the “pooled” charts will usually detect this type of problem more quickly.

Pooled charts are constructed by calculating a standardised Z score from the observed test result X. If the target for a given sample/batch is T and the (reproducibility) standard deviation is S, then Z = (X-T)/S. The Z values for the

pooled data can then be sorted into date order and plotted using the control chart techniques discussed previously.

The target for a pooled chart is always 0 and the control, warning and bias limits are normally drawn at ±K, ±W and ±B respectively, where K, W and B are the multipliers defined in section 1.8 above.

Whereas control charts for individual samples and the associated rules are mandatory for participating laboratories, pooled charts are usually created for information purposes only. Any additional mandatory charts should be specified in section 11 of the test procedure, along with any run rules that should be applied. Sometimes a point may violate a control limit on a pooled chart but not on the single sample chart due to rounding, or vice versa. In this case the single sample chart takes priority.

1.12 Constructing Control Charts using data from all laboratories

The aggregated data from all laboratories shall be used to monitor the overall severity and precision of the test. This can be done from the test monitoring web site by grouping all laboratories together on the same chart. Such charts may show a single sample (original scale) or multiple samples (Z scale).

The trend line in cross-laboratory charts is normally calculated with the smoothing parameter λ set to 0.1, since more data are available. The industry bias limits are typically set at $K\sqrt{\frac{\lambda}{2-\lambda}}$, where K is the multiplier of the SD which determines the control limits. If there is evidence of a change in severity or precision then the chairman and SDG LO should be informed so that a deeper investigation can be carried out.

Figure 3 shows an example of such a chart. The graph shows that the overall severity has remained constant. However it does appear that the later results contain fewer results outside the control limits, indicative perhaps of a slight improvement in precision.

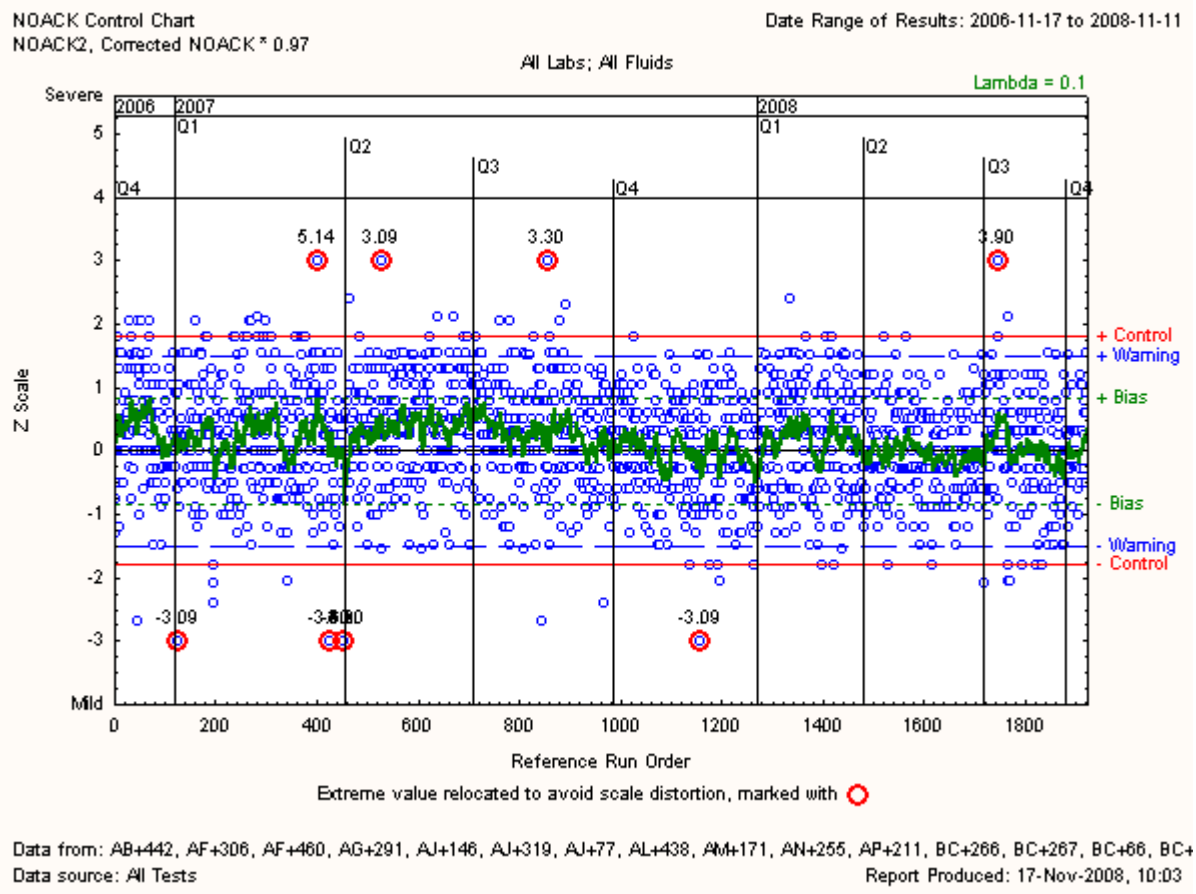

Figure 3. Example of a control chart combining data across laboratories to monitor the overall severity and precision

Since it is necessary to physically review these graphs on the test monitoring web site, responsibility for this action should be assigned by the chairman, typically to the DBA (database administrator). Such a review should be carried out every 1 to 3 months, commensurate with the number of tests being carried out.

1.13 Control Charting for New Test Methods, Reference Fluids and Batches

The first data to be plotted on control charts for a new method should normally be the round robin results which were used to support the approval of the test. If a laboratory/instrument did not participate in the round robin, then its control chart should start with the data used to qualify it as a new installation, as described in section 1.14. These initial control charts can then be used to determine whether a particular laboratory was in a state of statistical control at any given point in time, both before and after the test method was approved.

Careful thought needs to be given to candidate tests conducted (a) in the early stages of test development and/or (b) shortly after round robins, including studies conducted to assess new reference fluid batches.

CEC Guideline 18 allows candidate tests run before a test method is approved to become a valid CEC test, provided that the laboratory was in statistical control at the time. This is determined by applying the laboratory acceptance rules retrospectively. If there are subsequent reference tests, then these should also be plotted on a control chart and the usual control chart limits applied.

Laboratories may experience difficulties knowing whether they are in control when new reference fluid batches are introduced with unknown performance levels. Working groups should therefore consider performing pilot tests at a small number of laboratories in order to set provisional limits. The process of setting these limits should take into account (a) the increased levels of uncertainty in targets estimated from small pilot studies and (b) any historical data on batch-to-batch variation. Procedure 3 Appendix B gives guidelines for assessing the acceptability of round robin results.

Candidate tests will not be considered valid if the last reference or round robin result(s) before the candidate test are outside the acceptance limits, or if the number of candidate tests or elapsed time since the last reference exceeds the permitted maxima.

Working groups are encouraged to ensure that new fluid batches are prepared and tested before the previous batch runs out.

1.14 New Installations

The normal requirement for a new laboratory starting to run candidate tests is to test two reference samples. Where possible, these should be on a high performance sample and a low performance sample to demonstrate discrimination. The number of repeats will be that carried out by peer laboratories in the most recent full round robin. However, ideally, at least two repeats should be run on each sample.

If the requirement is for a single test on each sample, then it is recommended that the procedure require that both should be within the warning limits. If the laboratory has run more than one test on a sample, then only the most recent result on that sample should be considered. If one test is outside the warning limits then it should be repeated and the repeat should be within the warning limits. For high cost tests, this requirement may be loosened to allow one result between the control and warning limits, without the need for a repeat.

If the requirement is for 2 tests on each product then only the 2 most recent tests on each fluid should be considered. None of these must be beyond the control limits, and it is recommended that in total no more than 1 test result be outside the warning limits.

Some loosening of the limits may be permitted for tests with multiple parameters to allow for the increased risk of excursions outside the warning limits due to random variation (see Section 1.8).

For a laboratory that already has one or more stands in the referencing system, the requirements for adding additional stands may be reduced if appropriate. The minimum requirement is one test on each sample. Reference tests should also be conducted when there is a major hardware change, for example when a new test engine is installed on an existing stand in non-destructive fuels tests. In circumstances where the hardware has to be replaced frequently, one test may have to suffice in which case the result should be inside the warning limits.

Working groups should set up referencing requirements bearing in mind the general expectation is that once a laboratory is well established and in control, roughly 10% of its tests should be references.

1.15 Writing Test Monitoring into the Test Method

The test monitoring protocol for a particular test should be included in section 11 “Referencing and Precision Statement” of the test procedure. It needs to address the following items:

- Which test parameters are to be monitored (section 1.3)
- Which reference fluids and batches are to be tested (section 1.4)
- Frequency and order of reference testing (section 1.5)
- The Target, Control, Warning and Bias Limits for each fluid, or where this information may be found (sections 1.6, 1.8)
- The EWMA parameter λ (section 1.7)
- Any run rules being used (section 1.9)
- How to respond to violations of the limits (section 1.10)
- How new laboratories or stands can be brought into the referencing system (section 1.14)
- How often reference data should be uploaded to the database (section 3.8)

2 Analysing Test Monitoring Data

2.1 Overview

From time to time (typically before WG meetings), the reference data for each reported parameter should be reviewed by the SDG LO. The methods used for analysing round robin data in Procedure 1 and the START program can be used, with some adaptations, as described in sections 2.2 to 2.5. The output from such an analysis is typically used to complete the WG progress report.

Some of the questions that test monitoring data can be used to address are:

- Is there evidence of a global shift in severity or precision over time?
- What is the estimated reproducibility of the test and does it meet the reproducibility target? Note that the estimate of repeatability from test monitoring data is less reliable.
- Are there significant differences between laboratories in terms of severity or precision which deserve further investigation?

- Are there any variables which could explain systematic differences in results between or within laboratories?
- Should the test monitoring limits be revised, either because of a severity or precision shift or because the initial limits were subject to uncertainty?

If the WG and SDG LO agree that the control limits should be changed then the DBA should inform the database focal point (DFP). Modifications to the limits should be made simultaneously to both the test monitoring database and the test procedure. In both cases it should be clear when the new limits came into effect. Tests started on the day of the update (or later), should use the new limits.

2.2 Selection of the data to be used for the Statistical Analysis

The group should define a suitable time window for the data to be used in the statistical analysis. This will typically cover the last 6-12 months, although a longer time period may be used for tests where relatively few reference results are generated.

Only operationally valid results should be included in the analysis. Valid results outside the control limits should be retained, since these could be useful indicators of possible severity shifts.

2.3 Detection and Removal of Outliers

The outlier detection techniques described in Procedure 1, which are based on ISO 5725, assume that each laboratory carries out the same number of tests on each sample (typically 2). This is not generally the case for test monitoring data. If there are 3 or less results for each laboratory/instrument × sample combination then the techniques of Procedure 1 may be used. Appendix C describes how Cochran's test for repeatability outliers may be generalised when the number of repeats is not the same at each laboratory.

If there are more than 3 results per laboratory/instrument × sample combination then the outlier detection techniques of Procedure 1 are inappropriate and other techniques must be used. Unlike round robin analysis, there are no hard and fast criteria prescribed, so a combination of graphical techniques, statistical analysis and engineering judgement must be used. Note that since these data sets are larger than those typically generated in round robin studies, the impact of including/excluding a single result can be much less.

A number of outlier techniques are available which may be applied iteratively.

1. <u>Detection of individual results as outliers.</u>
   It is recommended that a result which is more than 3 standard deviations from the mean is excluded. The simple mean and SD are calculated from the data from all laboratories excluding the point in

question. Particular care needs to be taken if the number of repeats varies significantly from laboratory to laboratory.

2. Detection of Laboratories as outliers
Control charts, START raw data plots and Laboratory Comparison Plots can all be useful ways of visualising the data to assess whether a laboratory is significantly out of line with its peers. Grubb's test can also be used to compare laboratory means, but this is only approximate when the number of repeats varies. Periods where a laboratory is clearly out of control, or is trying to get back into control should be excluded.

2.4 Calculation of Means

The mean values for each sample at each laboratory can be calculated in the usual way. However, statistical advice should be sought when calculating standard errors and confidence limits as successive measurements at any particular laboratory are likely to be autocorrelated, and hence the usual assumptions of independence may be invalid.

The industry mean for any particular sample should be calculated by averaging the means at the various laboratories giving each laboratory equal weight, irrespective of the number of measurements taken. However, care needs to be taken before including laboratories with very small numbers of measurements. If laboratories have multiple stands then these may be considered as separate laboratories for the purpose of these calculations.

2.5 Reproducibility, Site Precision and Repeatability

Once the outliers have been identified and removed as described above, reproducibility and site precision (see definition below) can be calculated using section 7.4.5 of ISO 5275-2 and the START program. Reproducibility estimates obtained from test monitoring data are usually reliable.

Test monitoring data are typically collected over extended periods of time, perhaps by different operators, with several intervening tests on candidate fluids. Such conditions are more closely aligned to site precision conditions, as defined in ASTM D6299, than to repeatability conditions.[1] Site precision

[1] Repeatability r: The value equal to or below which the absolute difference between two single test results, obtained in the normal and correct operation of the same test method on identical material, may be expected to lie with a probability of 95% when conducted under the following conditions: non-consecutive tests with intervening changes of test material, completed in a short time interval by the same operator at the same laboratory using the same apparatus.

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

can be thought of as “long term” repeatability. Therefore the precision analysis given by a formal application of ISO 5725, ignoring the order of results from the same laboratory, gives estimates of site precision rather than repeatability.

Notwithstanding the above, test monitoring data can be used to get a simple estimate of short term repeatability from pairs of consecutive reference results using the following formula:

$$r_{cons} = 2.8\sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\left(X_{i+1} - X_i\right)^2}{2}}$$

This calculation is available as an option in START. However, since there can typically be up to 9 candidate tests between references, this estimate might be expected to over-estimate the (short-term) repeatability. Note that for n≥3, the degrees of freedom associated with this quantity will be less than n-1 as the quantities in the summation are not statistically independent.

Short term repeatability provides a useful yardstick for comparing pairs of fluids which are tested back to back, for example in “relative to reference” specifications. Site precision should be used for comparing results from the same laboratory collected over longer time frames or by different operators.

Due to autocorrelation between consecutive results at the same laboratory, statistical advice again needs to be taken when calculating confidence intervals for repeatability, site precision and reproducibility.

2.6 Laboratory/Instrument Comparison

Test monitoring data can be used to compare the relative performance of the participating laboratories.

One option is to use the Laboratory Comparison chart produced by the START program as described in Procedure 1 (Appendix C). It should be noted that the repeatability component of variation in such plots corresponds to site precision rather than short-term repeatability.

An alternative approach is simple graphical plotting. For each laboratory, we can calculate the mean, site precision SD and consecutive results SD, and draw graphs to compare these quantities. These graphs can be used to highlight laboratories with particular problems. For example one can plot the site precision standard deviation (and/or the consecutive results SD) versus the mean at the various laboratories for a single reference sample.

Figure 4 shows how such a graph might look. The mean at most laboratories is greater than the target, which indicates that the target and limits should be reviewed. Clearly laboratory Q is very low compared to the other laboratories, and is likely to be outside the new control limits. Laboratory Q and possibly the working group should consider the likely causes. The standard deviations

all lie in the same band, so none of the laboratories have serious repeatability issues.

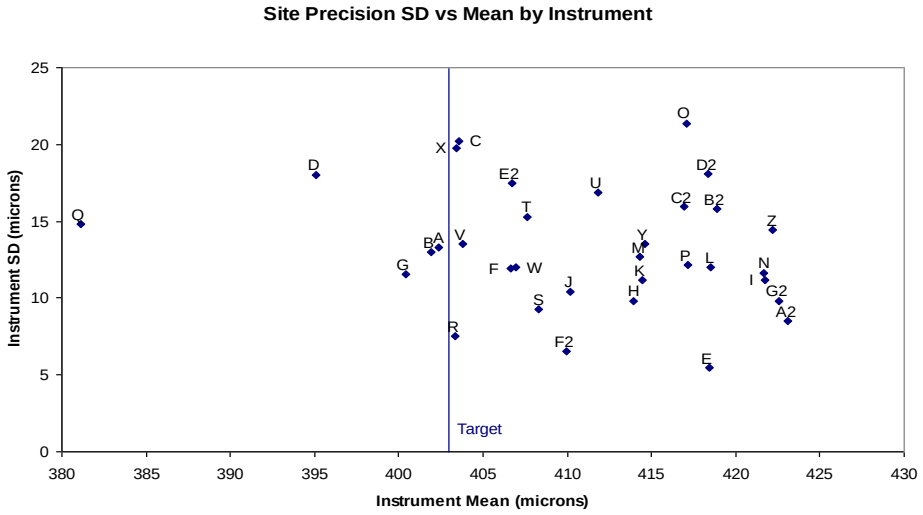

Figure 4. Site precision SD vs. Mean Chart for the Wear Scar Diameter from 33 HFRR instruments

2.7 Comparison with Previous Time Periods

Plots can be constructed comparing means and standard deviations with earlier time periods and round robins. Detailed discussion and an example are given in section 9 of Procedure 1.

As noted previously, the autocorrelation of successive results can complicate the calculation of meaningful confidence intervals. This also complicates the comparison of means and precision estimates from different years.

Note also that it is incorrect to use standard F-tests to compare (squared) reproducibility or laboratory-to-laboratory variability estimates from different years. These will usually be estimated from results from roughly the same set of laboratories and so be non-independent. This issue also affects the comparison of means. The calculation in START assumes independence. This gives a potentially weaker test for detecting a difference than if the laboratories had been taken into account. Statistical advice should be taken.

2.8 Revision of Test Monitoring Limits

Test Monitoring targets and limits should be reviewed at the end of the analysis and changes should be considered either if there has been a change in severity or precision, or if it is considered that including the new data would give improved estimates. Options are (a) to leave the limits as they are; (b) to update the limits using the new data in conjunction with some or all of the old data; (c) to calculate new limits using only the new data.

Option (a) might be chosen if there is little evidence of change in severity or precision and/or if the current test monitoring 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 limits are based on a relatively small data set, and if there is no evidence of major change in the new data. Option (c) would only be chosen if there is clear evidence of a major change in severity and/or precision and this is based on a substantial data set. If the impact of any change would be considered minor then option (a) should be preferred. Frequent changes in limits are not desirable. Further issues which should be considered are (a) how many laboratories and how much data the current targets and limits are 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 to the Management Board and noted in the progress report. This may also need to be reported to the setters of any specifications based on the test.

2.9 Additional Analyses

Statistical techniques such as Analysis of Covariance can be used to study the effects of factors such as ambient temperature and/or pressure on the test result. This might help explain systematic variations between and within laboratories.

3 Data Depositories

3.1 Introduction

In order to carry out test monitoring and the subsequent statistical analysis effectively, it is necessary that the reference data from the test laboratories be collected and stored efficiently and reliably. This is managed by creating a “Data Depository” for each test. Even for test methods where formal test monitoring is not being carried out, the data depository has value in allowing a working group to collect test data in a consistent format and subsequently produce data summaries and perform statistical analyses.

A data depository is a central web-based electronic database for the storage and retrieval of CEC reference data. It provides a permanent archive for data generated by the working group. It is also the basis for the on-line test monitoring system. The data dictionary, developed by the working group, defines the data which will be collected and the format in which it will be stored.

CEC reference results are stored in two separate data depositories. The ATC-ERC database covers methods used in ACEA engine tests for lubricants; both reference and candidate results are stored. The remaining test methods are stored in the CEC-TMA database where only reference results are kept.

Responsibilities for maintaining the electronic test monitoring systems are detailed in Appendix A.

3.2 Applicability of Data Depositories to CEC Tests

A data depository should be created for a test method wherever there is value in storing test monitoring, round robin or calibration data for further analysis. It can also reduce the effort required for data acquisition, storage and retrieval.

Working groups may be granted exemptions in rare cases if it can be demonstrated that there is no value in keeping historical raw data in a unified format. Working groups which only run round robins should normally keep a data depository in order to provide secure storage for their data and to compare precision and laboratories from year to year.

3.3 The Data Dictionary

The information stored in a data depository is defined in the data dictionary.

Appendix B shows two examples. The first is for the Kurt Orbahn shear stability test in the CEC-TMS database. The second example is the dictionary for the Peugeot TU572 Viscosity Increase and Piston Merits test, held within the ATC-ERC database.

Both data depositories have a similar structure as reflected in their data dictionaries. The ATC-ERC data depository uses the same data dictionary for both reference and candidate data and thus has more entries.

The parameters in yellow are standard fields which will be present for most or all test methods. These include identifiers for the laboratory, instrument, fluid, batch, date/time, test number, operator, validity, keyword and data source (round robin or test monitoring). These fields are likely to be very similar for all test methods although, as noted previously, this list is far more extensive for the ATC-ERC database.

The parameters in green are the key test results which are required to be reported by the test procedure.

The parameters in orange define other pieces of information that the group wish to collect specific to the test, e.g. run time parameters or ambient conditions or supplementary test result parameters. A key part of drawing up the data dictionary is deciding how much additional data is worth collecting.

Both examples have a number of these supplementary parameters. These include:

- *Intermediate Parameters* used to calculate the reportable parameters. For example, for the Kurt Orbahn shear stability test, KV_INIT is used in the calculation of the rateable parameter SHEAR. And in the TU572 test, there is a rateable parameter PC_AV_M5 which is determined as the average of 5 ratings. These individual ratings are recorded.

- *Operational Data*
  For the Kurt Orbahn test, pressure which is set during calibration and temperature during calibration, which is not.
  For the TU572 test, the mean power output at phase 1.
  Note: For key operational parameters which vary throughout the course of a test, it is recommended to record summary statistics such as the mean and the standard deviation. If the Working Group has defined Quality Indices to rate the operational data, these should also be included.
- *Identifiers* for pieces of hardware which are re-used in subsequent tests, e.g. NOZZL_ID in the Kurt Orbahn test.

Some CEC tests, particularly engine tests, produce large amounts of operational data and it is not practical to store all of this in the data depository. The Working Group must decide which measurements have long term value when designing the data dictionary.

Data dictionaries may be changed from time to time if, for example, the CEC working group wishes to study additional parameters. If there is a major change to the method then either the version of the method needs to be included as a parameter, or a new set of output parameters needs to be defined. The second method prevents pre- and post-change data appearing on the same control chart. For example, the NOACK group introduced a 0.97 correction factor in 2007. Initially uncorrected and corrected test results were stored in separate variables NOACK and NOACK2. Subsequently the storage of uncorrected test results (NOACK) was discontinued.

3.4 Setting up a Data Depository

CEC Working Groups should start to use a data depository as soon as a test method becomes well defined, to ensure that all the reference data are collected and validated. Typically this would be the point at which Surveillance Group (SG) status is achieved, when the key outputs are well defined, and no major changes are envisaged. It is highly desirable that the data from the round robin which leads to the group being granted SG status is captured. However it may be necessary to store these data retrospectively.

For tests not covered by the ATC-ERC database, a project to set up a data depository for existing working groups will be run by the CEC Secretariat. The key technical activities are:

- Determining the Data Dictionary – and getting agreement from the working group. The SDG representative should advise on what data has been useful in past analyses and what might be used in the future if available. For example, if the overall rating is the average of a number of individual ratings and previously statistical analyses have examined the individual ratings to look for patterns or outliers, then it may be useful to retain the detailed rating data.

- Compiling the historic data for entry into the data depository. This task may be carried out by the Database Administrator with the advice of the SDG LO.
- Setting up a test input sheet to enter data. This should be designed by the DBA and constructed by the DFP, based on the group's requirements. The SDG LO should review the sheet, to ensure that there is a minimal risk of data being entered incorrectly.

3.5 Elements of a Data Depository

The key elements of a Data Depository are Data Entry, Validation, Upload, Storage, Security and Reporting. These are described in detail in sections 3.6-3.8 below. The requirements described are for the CEC-TMS system only.

3.6 Data Entry

The system should be designed to minimize the effort required to enter the data, and help protect against data entry errors. If the hardware used for running the test can create the upload file, or part of it, then that is preferred. However for most of the bench tests the data will need to be manually entered.

For the CEC-TMS system, data is entered into a "Test Monitoring Input Sheet". This workbook has a number of "data entry" worksheets, which are specially designed for each individual test. The workbook takes data from the "data entry" forms and generates a flat file which is subsequently uploaded to the CEC-TMS database. The Test Monitoring Input sheet should also provide initial validation of the input data to reduce the risk of incorrect data being uploaded to the database.

When designing the data entry forms for the Test Monitoring Input Sheet, the following points should be borne in mind:

- Clear Layout -This sheet should be well laid out so that the user can easily identify where each piece of data should go.
- Clearly identifying the required input format - where ambiguous, the sheet should make clear in exactly what format or units the data are required. For example, the number of digits after the decimal separator should be specified. Particular care needs to be taken with date/time fields as different conventions are used in different countries.
- Provision of a list of possible inputs – where a field can only take a set number of values, provide these to the user in a drop down list.
- Data which should be re-entered every test should be cleared from the data entry area once the previous test has been stored. This will include all the test results and operational variables, as well as the date/time stamp. Where convenient these fields should be stored in the same part of the data entry sheet.
- Data which stays the same for each test (e.g. the laboratory name), should not need to be re-entered for each test run, and may where convenient be put into a different area.

- A clear help system.
- Minimize the number of fields that have to be entered - e.g. calculate parameters where possible.

If the data entry sheet is updated, then this should be immediately communicated to all individuals entitled to upload data.

3.7 Data Validation

Poor quality data, arising from data being input incorrectly, can severely compromise control charts, statistical analyses and other reports from the data depository. Validation by the data depository should work at two levels. Firstly the user should be warned if entered values are outside the typical range and need to be checked. For example, if a control chart parameter is outside the warning or control limits, then this should be flagged straightaway. If the date entered is very different from the current date this should also be flagged. Secondly, the system should reject any data which is definitely incorrect. For example, rejecting atmospheric pressures outside the range 0.9 to 1.1 bars would flag results measured in the wrong units e.g. millibars or mmHg.

For usability reasons, it is preferred that the checks are carried out as soon as possible after the data have been entered, and where possible before the data is uploaded. However some checks e.g. on the laboratory name, instrument and reference sample ID, have to be carried out after the data are uploaded.

3.8 Uploading and Storage

There should be a facility for nominated individuals to remotely upload data to the database via the internet. Once uploaded, a laboratory should be able to review that data. If any errors are found then it should be possible for the laboratory to make the necessary corrections, but the database should record that a correction has been made.

Laboratories should upload their data as soon as possible after the completion of each reference test, in order that industry trends can be monitored. In the interests of laboratory efficiency, if the reference tests are conducted on a frequent basis, laboratories may submit a batch of results, provided that results are always submitted within 3 months of the completion of the test. In such circumstances laboratories must check each test result against the control and warning limits as soon as the test is complete. Laboratories should submit data immediately if their EWMA line is approaching the bias limits and the new result is liable to lead to an excursion outside the limits.

When a working group meeting is due, the DBA should request laboratories to submit their reference data by a specified date in advance of the meeting so that the SDG LO can carry out a precision and trend analysis.

Test monitoring, round robin and calibration data, should all be recorded in a data depository, although the different types should be clearly distinguished. Candidate data is not required for test monitoring. (Candidate data is uploaded to the ATC-ERC database but not to the CEC-TMS database.) Data from invalid data tests should be recorded. These results should be clearly indicated, since these will normally be excluded from statistical analyses and test monitoring plots. The reason for the invalid test should be recorded using an appropriate keyword.

Test results outside the control limits are not necessarily invalid. If, however, a test result outside the control limits was entered as valid, but subsequently found to be invalid, then that result should be recoded as invalid, and the cause of the problem should be recorded.

It is the responsibility of the Working Group to set limits for operating parameters, and define the requirements for a valid result in the test procedure.

3.9 Security

The system must only allow nominated individuals from each company to upload data.

Additionally, only employees of companies which are members of the relevant CEC Working Group should be able to view control charts or data from that group. These reports should code the identity of the laboratories. Laboratories should be able to identify only their own results and not those of other laboratories.

Only the chairman, SDG LO and DBA should have access to reports which contain the laboratory codes, or otherwise reveal the identity of laboratories.

3.10 Reporting

The data depository should, on demand, be able to produce control charts using the methodology described in section 1. It should be possible to include all the data that has been uploaded, or some subset thereof. The control charts for individual laboratories or stands should allow the user to easily determine whether the test is currently in statistical control. Additionally, it should be possible to produce control charts covering all laboratories as described in section 1.12.

Additional graphics may also be helpful. These include:

- Control charts based on Z-scores for multiple batches or multiple samples, as described in section 1.11.
- XY scatter plots between pairs of variables in the data dictionary.
- Discrimination plots which allow comparison of different batches of the same reference fluid, as exemplified in Figure 5. This plot shows that the results for batches 3 and 4 of reference fluid RL223 have had a

constant mean. There is however evidence that batches 4 and 5 of reference fluid RL172 gave results with different means.

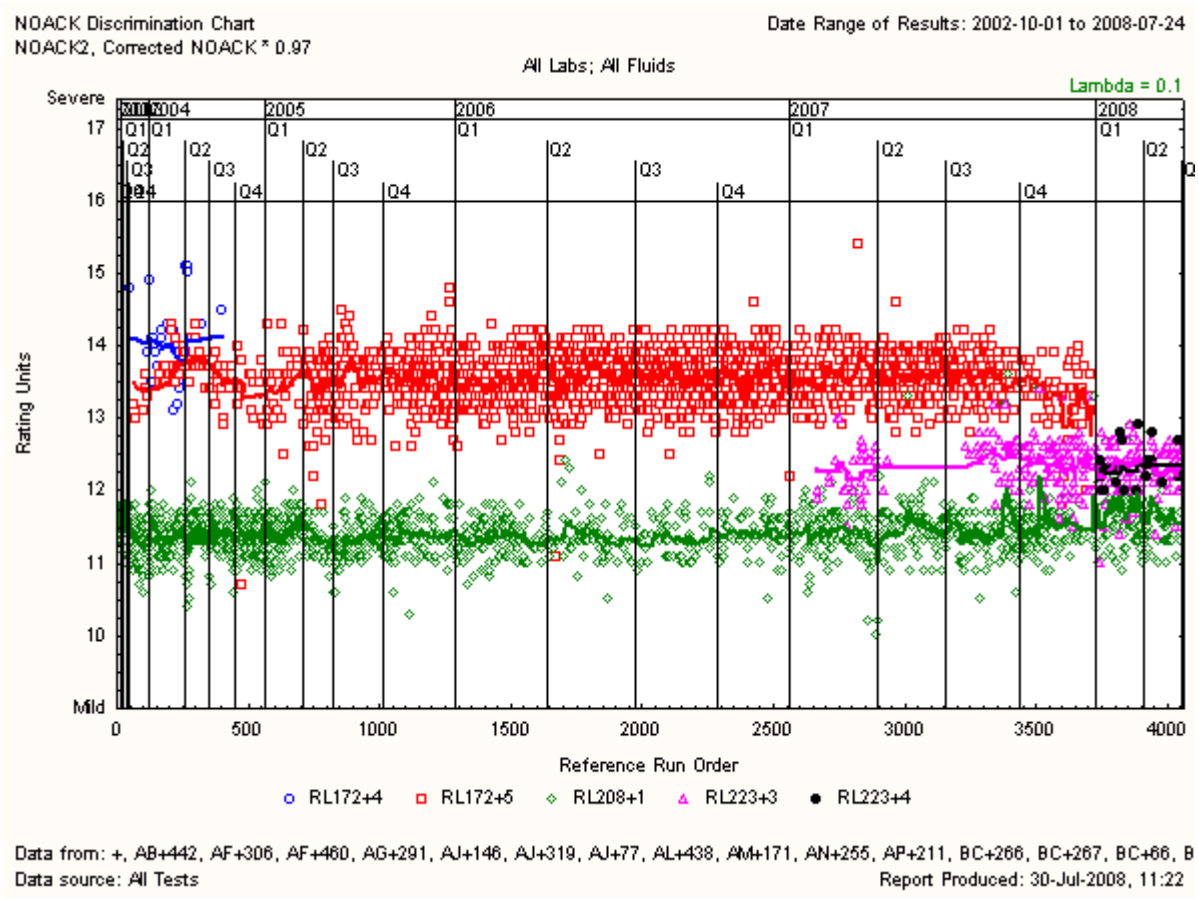

Figure 5. Example of a Discrimination Plot

It should be possible to download data from the data depository to an EXCEL spreadsheet. This format should be consistent with the requirements for the SDG statistical analysis program START. It should also be possible to download targets, standard deviations, control, warning and bias limits and their dates of application. Also it should be possible to download the data behind any graph.

Appendix A: Responsibilities for maintaining the test monitoring system

The responsibilities of the various parties are:

WG Chairman

- Ensure that the roles and responsibilities in the WG are filled. These include:
  - The database administrator (see section below)
  - Any responsibilities for "policing" laboratories' participation in the test monitoring system
  - Responsibilities for ensuring that industry trends are examined on a regular basis between group meetings – typically given to the DBA
- Provide advice (or delegate the responsibility for providing advice) to laboratories which may have difficulty staying within the limits.
- Pass concerns from group members about possible trends in the reference data on to the SDG LO for further investigation

SDG Liaison Officer (LO)

- Carry out periodic analyses to monitor the severity and precision of the test method as described in section 2.8, and make any recommendations for changes to the precision statement.
- Report the results of the analysis to the working group and alert the chairman to any evidence of changes in severity or precision.
- Provide information suitable for inclusion in the progress report for the test
- In consultation with the working group, determine Control, Bias and Warning limits, based on guidance from the working group
- Ensure that new limits are communicated to the CEC-TMS or ATC-ERC Database Focal Point (DFP) as appropriate
- Respond to concerns, passed on by the chairman, about possible trends in the reference data

Database Administrator

- The DBA is the primary contact between the working group and the DFP. The DBA should inform the DFP of any pertinent changes including:
  - Any new reference samples, new batches, new laboratories, or laboratories which are leaving the system.
  - For round robins, any new samples that are being used and also an annotation to show that the data are for a specific round robin and not general test monitoring results.
  - Changes in targets or limits.
  - Any significant changes to the test method, which may lead to shifts in severity or precision and which may require changes to the data dictionary or some other part of the test monitoring system.

- Review Industry Control Charts for changes in severity and precision on a regular basis – commensurate with the frequency of testing, typically every 1 to 3 months, depending on the number of reference results being generated.

Test Laboratories

- Carry out regular reference tests
- Check each new result against the test monitoring rules in force, and take appropriate remedial action if necessary
- Upload reference tests as required (see section 3.8)
- Advise the DBA and the DFP of any new stands/instruments which are joining the system
- For ATC/ERC registered tests, each laboratory will identify their reference test results by signing and returning the Reference Test Acceptance and Identification Form. This is available as Proforma 3 on the CEC website.

Database Focal Point (DFP)

- Maintain the data depository, ensuring it can be accessed by CEC members
- Provide laboratories with logon IDs and passwords
- Respond to any technical difficulties with the system
- Respond to requests for changes from the DBA, within the remit agreed in the contract with CEC. This includes the Test Monitoring Input spreadsheet.

Appendix B: Examples of Data Dictionaries

Example 1: Kurt Orbahn test from the CEC-TMS database

Seq	TestType	Name	Len	Dec	Type	Unit	Description
10	KO	TESTTYPE	6		C		Test type designation. Must be exactly 'KRL'
20	KO	VERSION	8		C		Data dictionary version number (YYYYMMDD).
30	KO	ACKEMAIL	40		C		Email address for messages regarding EDT status.
40	KO	TEST_NUM	6	0	N		Incrementing test number for INSTR_ID. This field is blank for new tests uploaded to ERC. ERC will generate the number and post it on the EDT webiste for each test. If a lab wishes to correct previously submitted data, then enter TEST_NUM assigned by ERC into this field with corrected test data, and ERC
50	KO	LAB_CODE	2		C		Laboratory Code, as assigned by ERC.
60	KO	LAB_NAME	40		C		Laboratory Name
70	KO	OIL_TYPE	12		C		Oil Reference eg X
80	KO	OIL_CODE	2		C		Oil Batch Code e.g. 1, 2, …
90	KO	INST_MOD	10		C		Instrument model / generation
100	KO	INSTR_ID	4		C		Instrument Serial number / ID
110	KO	CANDIDAT	3	0	N		No. of candidate tests run since last reference
120	KO	OPERATOR	20		C		Operator Name / ID
130	KO	DATETIME	16		DT		Test start date/time (dd/mm/yyyy hh:mm), hours 00 to 23
140	KO	VALIDITY	15		C		Assessment of operational validity of the test. Pick one value from designated list of values.
150	KO	COMMENT	70		C		Comments regarding test operation or validity, optional.
160	KO	KEYWORD	15		C		Keyword selected from finite list of designated words regarding test validity. A value is required for invalid tests.
170	KO	RNDROBIN	20		C		Round robin designation for industry sponsored round robin or other matrix testing. Leave blank for tests run as part of test monitoring.
180	KO	KV_FINAL	6	2	N	mm/s2	KV100 after shearing
190	KO	SHEAR	6	2	N	%	Percentage Shear Loss
200	KO	KV_INIT	6	2	N	mm/s2	KV100 prior to testing
220	KO	PRESSURE	3	0	N	Bar	Pressure
230	KO	TEMP10CY	2	0	N	°C	Oil temperature after 10 cycles
240	KO	NOZZL_ID			C		Nozzle Identifier

Notes

- Yellow fields are normally included in most data dictionaries. Some of these fields will be fixed for a particular laboratory and test.
- Green fields are the performance results for the test
- Orange fields are additional information that the working group has decided to collect
- "Len" is the maximum length of the field in characters
- "Dec" is the number of decimal places
- "Type" is C for character or text data; N for numerical data or DT for date time
- "Seq" does not need to be completed by the working group

For numerical fields, minimum and maximum values should also be supplied (in additional columns which are not shown above).

Example 2: Peugeot TU 572 Test in the ATC-ERC Database (extract)

Seq.	Form	Area	Name	Len.	Dec.	Type	Unit	Description	Comments/Examples
10	1	TU572	VERSION	8		C	YYYYMMDD	TU572 20050331	This is the version identifier of this data dictionary
20	1	TU572	ENGINE	15		C		Engine identification code	TU572 (must be exact ATC engine test type code)
30	1	TU572	TESTPROC	15		C		Test procedure designation	L-88-02 (Must be in this format)
60	1	TU572	SPONSOR	40		C		Sponsor	as requested by sponsor
140	1	TU572	LABNAME	40		C		Laboratory	as defined by laboratory
180	1	TU572	STAND	5		C		Laboratory test stand	as defined by laboratory
190	1	TU572	LABCONT	20		C		Contact Laboratory	as defined by laboratory
220	1	TU572	OILCODE	38		C		Sponsor oil Code	as defined by sponsor
510	1	TU572	FUELSUPP	40		C		Fuel Supplier	Name of company supplying fuel
520	1	TU572	FUELCODE	20		C		Fuel	DF-95-03
530	1	TU572	FUELBTCH	8		C		Fuel batch	e.g., 02; *variable B.12*
540	1	TU572	R1OILCOD	38		C		Reference oil code (RL216), incl. Batch number	e.g., RL216/02
550	1	TU572	R1FORM	38		C		ATC/ERC Formulation Stand code (RL216)	format per Code Form E.2 for reference
560	1	TU572	R1TESTNO	30		C		Laboratory reference test identification code (RL216)	as defined by laboratory
570	1	TU572	R2OILCOD	38		C		Reference oil code (RL194), incl. Batch number	e.g., RL194/05
580	1	TU572	R2FORM	38		C		ATC/ERC Formulation Stand code (RL194)	format per Code Form E.2 for reference
590	1	TU572	R2TESTNO	30		C		Laboratory reference test identification code (RL194)	as defined by laboratory
600	2	TU572	V40_IN_A	7	1	N	mm²/s	Absolute Viscosity Increase (max - min)	*variable A.1*
610	2	TU572	PC_AV_M5	4	1	N	Merit	Piston Merit 5 elements	*variable A.2*
620	2	TU572	OCSOTEOT	6	3	N	kg/test	Oil consumption	*variable A.3*
630	2	TU572	RS_P1_G1	5	1	N	Merit	Ring Sticking Piston 1 (first ring)	*variable A.4*
640	2	TU572	RS_P2_G1	5	1	N	Merit	Ring Sticking Piston 2 (first ring)	*variable A.5*
650	2	TU572	RS_P3_G1	5	1	N	Merit	Ring Sticking Piston 3 (first ring)	*variable A.6*
660	2	TU572	RS_P4_G1	5	1	N	Merit	Ring Sticking Piston 4 (first ring)	*variable A.7*
670	2	TU572	PC_AV_L2	4	1	N	Merit	Land 2, Piston Cleanliness Average all Pistons	*variable B.1*
680	2	TU572	PC_AV_L3	5	1	N	Merit	Land 3, Piston Cleanliness Average all Pistons	*variable B.2*
690	2	TU572	PC_AV_G1	4	1	N	Merit	Groove 1, Piston Cleanliness Average all Pistons	*variable B.3*
700	2	TU572	PC_AV_G2	4	1	N	Merit	Groove 2, Piston Cleanliness Average all Pistons	*variable B.4*
710	2	TU572	PC_AV_G3	5	1	N	Merit	Groove 3, Piston Cleanliness Average all Pistons	*variable B.5*
720	2	TU572	V40_SOT_	6	1	N	mm²/s	Viscosity at 40ºC, 0 hour	*variable B.6*
730	2	TU572	V40_012_	6	1	N	mm²/s	Viscosity at 40ºC, 12 hours	*variable B.7*
740	2	TU572	V40_024_	6	1	N	mm²/s	Viscosity at 40ºC, 24 hours	*variable B.8*
750	2	TU572	V40_048_	6	1	N	mm²/s	Viscosity at 40ºC, 48 hours	*variable B.9*
760	2	TU572	V40_060_	6	1	N	mm²/s	Viscosity at 40ºC, 60 hours	*variable B.10*
770	2	TU572	V40_072_	6	1	N	mm²/s	Viscosity at 40ºC, 72 hours	*variable B.11*
780	2	TU572	EP_T1_AV	5	1	N	kW	Mean Phase 1 Power Output	*variable B.13*
790	2	TU572	FC_T1_AV	6	2	N	kg/h	Mean Phase 1 Fuel Consumption	*variable B.14*

Notes

- The same comments apply as for example 1, except that minimum and maximum values are not needed.
- “Form” does not need to be completed by the Working Group.

Appendix C: Generalisation of Cochran's Test to uneven Group Sizes

ISO 5725 part 2, section 2, paragraph 7.3.3.6 describes the use of Cochran's Test for detecting repeatability outliers, in a round robin. However the method, as written, only applies when all laboratories that have tested a sample more than once, tested it the same number of times. This section explains how the method may be generalised to arbitrary numbers of repeats.

For a given reference sample, suppose that Laboratory i has n_i repeat tests, with a calculated standard deviation of s_i, based on an (n-1) divisor. Let the corresponding degrees of freedom n_i-1 be denoted by ν_i. Laboratories which have not carried out at least two tests are excluded from the analysis since they provide no repeatability information. We make the same statistical assumptions as ISO 5725 that all results are normally distributed, with a common within laboratory standard deviation, and that within a laboratory, all results are independent.

Let $\nu = \sum_i \nu_i$ be the pooled degrees of freedom for the sample.

Under the null hypothesis that the standard deviation of each laboratory is equal, the distribution of:

$$F_i = \frac{{s_i}^2}{\left(\sum_{j \neq i} \frac{\nu_j}{\nu} {s_j}^2 \right)} \; has\ an\ F-distribution\ on \left[\nu_i, \nu - \nu_i \right] \text{degrees of freedom}$$

From this F-statistic a p-value can be obtained for each lab, based on a one-sided test.

The p-values for each lab can be combined into a single p-value by taking the smallest value and multiplying it by the number of comparisons being carried out (equal to the number of laboratories). This is known as a Bonferroni correction. This overall p-value is compared to the standard significance limits of 0.01 and 0.05 to determine whether the result is an outlier or a straggler.