Test Monitoring Center
6555 Penn Avenue
Pittsburgh, PA 15206-4489
(412) 365-1000

MEMORANDUM: 02-105

DATE: October 29, 2002

TO: Single Cylinder Diesel Surveillance Panel

FROM: Scott Parke

SUBJECT: Quality Index

Among the topics on the agenda for our upcoming meeting on November 7 is Quality Index. In the attached document, I've put together a description of the process used to generate the various components of a Quality Index system. I realize that it is lengthy (11 pages) but I think that our discussion at the meeting will be more productive if everyone involved is familiar with the material in this document.

SDP/sdp/ m02-105.sdp.doc
c: F. M. Farber
ftp://tmc.astm.cmri.cmu.edu/docs/diesel/scote/memos/mem02-105.sdp.pdf

distribution: Email

Why QI?

In all forms of testing one goal overrides all others: obtaining repeatable results. Whether the field is medical testing, scholastic achievement testing, or stationary engine testing repeatability is paramount. A test that can't reproduce the same results, time after time, under the same conditions is no test at all.

It is reasonable to expect that doing the same thing under the same conditions will yield the same results. Note that a key part of that statement is "under the same conditions". Until it becomes possible to rewind time, one condition of "under the same conditions" is already broken every time a test is run after the first. Just how much more room for interpretation is there in "under the same conditions"? This article will explain how the lubricant testing industry uses a measurement called Quality Index to determine just how close to "under the same conditions" each test is.

Perhaps the most obvious way to test the performance of an engine oil is to pour it into an engine and run it. Repeat the process several times for different oils and you will be able to compare the performance of the oils. Unfortunately, a running engine is an extremely complex system with hundreds of factors influencing oil performance (things such as speed, temperatures, pressures, flow rates, etc. collectively referred to as operating conditions). Is the performance difference exhibited between two oils due to the composition of the oils, or is it due to one engine being run 1000 r/m slower than the other?

Of course a test developer is smart enough to stipulate that all tests be run with the engine spinning at, say, 3000 r/m. But in order to test the particular lubricating property of interest it is usually necessary to run the engine continuously for a considerable length of time – usually for days and in some cases for as long as a month. It is not economical nor, indeed, even possible to control the speed of an engine to *exactly* 3000 r/m for a period of 250 or 500 or more hours.

fig. a1 Ideal case of all data exactly on target

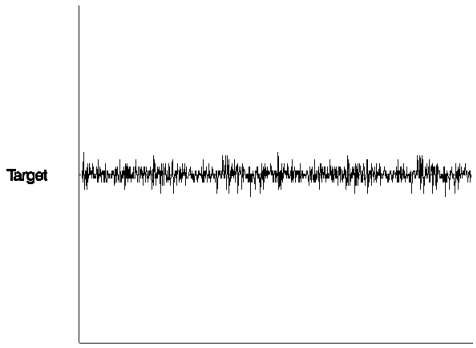


fig. a2 More realistic case showing data fluctuation

In practice, all specifications have a tolerance. The desire for repeatability drives tolerance lower; economics drives tolerance higher. A test developer must balance these two competing interests when specifying how closely operating conditions must match the targeted value.

What measure should be used to determine whether or not the operating conditions of any particular test matched the targeted value? The operating conditions of an engine can be recorded hundreds of times a minute. One measure would be to require that all of this data fall within a given tolerance.

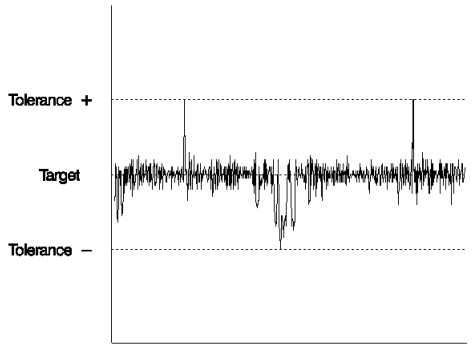


fig. b Tolerance set at minimum and maximum data values

In order to ensure that *all* data falls within the tolerance, however, either the control system must be very precise or the tolerance must be set wide enough to account for any variability. As already stated, increased control precision may not be economically feasible. This leaves widening the tolerance as the only available option. The graphs below show why this approach is undesirable.

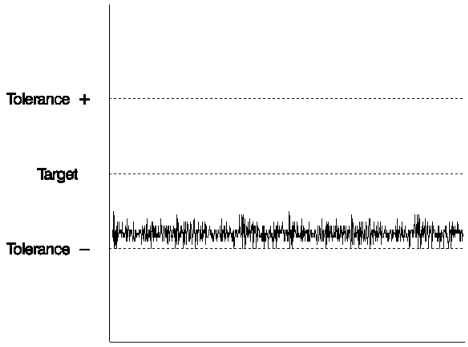


fig. c1 Data is precise but off-target

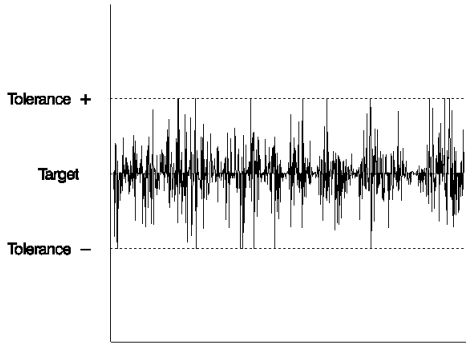


fig. c2 Data is on-target but imprecise

All data in both of these examples is within the tolerance but a test run with one or two or even most of the data points at the targeted value clearly is not the same as one run with all of the data there.

One measure that has been used to address this concern is to require the average of all data points to be within a given tolerance. With the data averaged, the tolerance can be set much tighter than if each individual data point is required to be within tolerance.

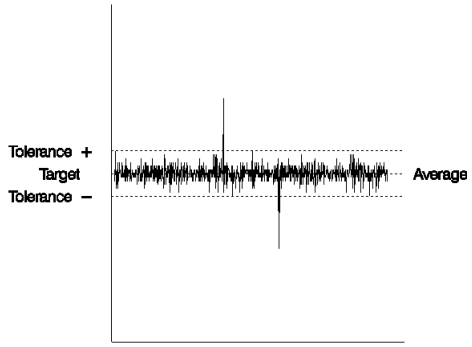


fig. d Average of data within tighter tolerance

The graph below shows one of the flaws in this approach.

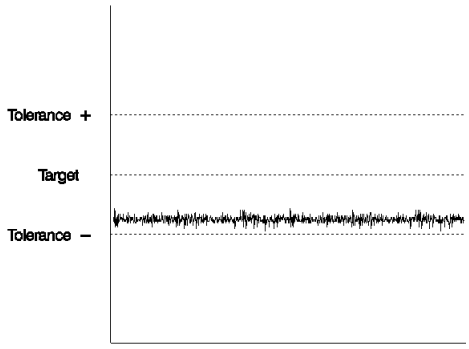


fig. e Average is within tolerance but not on-target

This graph shows another.

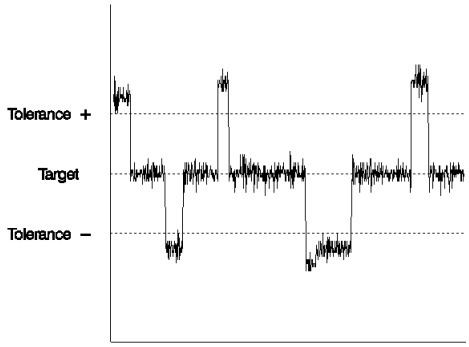


fig. f Average within tolerance and on-target

In the first case, the average of all data (and, in fact, *each and every* data point) is within tolerance but is not at the targeted value. In the second case the average is likewise within tolerance *and on target*, but numerous excursions beyond the allowable tolerance occurred.

A better controlled test will "consume" less of the allowable tolerance than one with poorer control. In lubricant testing, a measure called offset percent has often been used to quantify the amount of tolerance consumed. As the name implies, offset percent is the difference between the average of the data recorded and the targeted value expressed as a percentage of the allowable tolerance. For example, if a target is 100 with a tolerance of +/-10 then a test run at an average of 98 would have an offset percent of 10%.

$$\%off = \frac{|98-100|}{20} \times 100\% = 10\%$$

Similarly, a better controlled test will experience fewer and smaller excursions beyond the allowable tolerance than one with poorer control. A measure has been devised to quantify this aspect of test control as well; it is called deviation percent (or sometimes %out – out being short for outlier) and is the summation of the percentage of the data that was recorded outside the tolerance. Again using a target of 100 and a tolerance of +/-10, a data point recorded at 112 on a 100 hour test recording data once every 6 minutes (0.1 hours) would contribute 0.02% to the total deviation percent.

$$\%out = \frac{|112-110|}{10} \times \frac{0.1hours}{100hours} \times 100\% = 0.02\%$$

Summing this value calculated for every point recorded outside the allowable tolerance will give the deviation percent for the entire test.

Consider the situation presented in this graph, however:

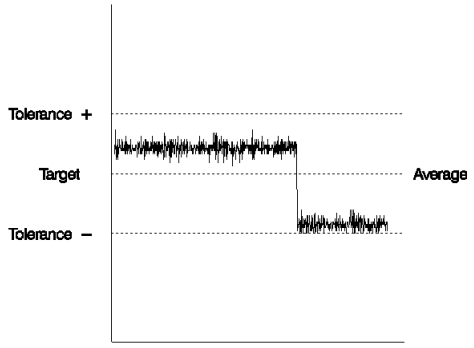


fig. g Offset = 0%; deviation = 0%

The test average is exactly on target. The offset percent is 0%. The deviation percent is 0%. Obviously, this test is not the same as if the test had held to the targeted value for its entire length. Running an engine 50°C hotter than it should be can not be compensated for by running 50°C cooler for an equal time period (or 25°C cooler for twice as long, etc.).

In manufacturing, products must be evaluated to determine whether or not they are fit for use. For example, a part can be measured to ensure that it is neither too long nor too short. Knowing that all of the parts in a batch of hundreds are within a specification tolerance does not really allow for comparison of one batch to the next. All that can be said is that they were all within tolerance; one batch may have all been on the short side another may have been randomly scattered all throughout the range and still another may be all exactly at the specified length. Can the three different machines that produced these three batches all be said to perform equally well?

When considering a collection of data, knowing the average value of the data is more helpful if it is known how far from that average any single point is likely to be. The difference from the average value to each point can be determined. The standard deviation for a collection of data is an expression of the average value of these differences. The smaller the standard deviation, the more tightly grouped is the data.

Unfortunately for use as a device for monitoring the performance of the machines in the previous manufacturing example, standard deviation only tells how data points are distributed in relation to each other, not in relation to where they ought to be. Thus, while the standard deviation on the second of the example machines will be poor, the first and third will both be equally good.

So, in order to adequately monitor the performance of the machines in the example it is necessary to know the following: the targeted length, the average length of the parts actually produced, the tolerance in the targeted length, the standard deviation of the lengths produced, and the differences between all these

values in relation to each other. Knowing *all* of this will indicate that the third machine's performance is the highest.

Complicating matters, however, is that a given part is likely to be subject to not one screening criteria or parameter (length in the above example) but several. Keeping track of different targets, tolerances, averages and standard deviations and the acceptable limits on each becomes cumbersome.

Here a review of the desirable characteristics for a measure evaluating the output of a machine in a production environment is in order. The measure should:

1) describe how close the parts are to the target for each parameter
2) describe how different the parts are from one another for each parameter
3) allow for easy comparison across parameters irrespective of units of measure or the magnitude thereof
4) allow for batch-to-batch comparison regardless of batch size

Such a measure, known as "Quality Index" or QI, exists (see the September, 1984, issue of Evaluation Engineering for derivation) and is defined as follows:

$$QI = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{U + L - 2X_i}{U - L}\right)^2$$

Where: X_i = measurement being made
n = number of measurements made
U = allowable upper limit of X (top of tolerance)
L = allowable lower limit of X (bottom of tolerance)

When all values for X are equal to the targeted value QI will evaluate to 1; when X is equal to either U or L QI will evaluate to 0. X values falling within the range of the tolerance will influence the overall QI number up or down between 0 and 1 depending on how close they are to target.

Note that the QI calculation is very similar to a standard deviation calculation but differs in two important ways. First, the standard deviation describes how the X values compare to each other while QI compares each X to the target value. And, second, QI is both unitized and dimensionless. It is this second aspect of QI that allows it to be used to compare any of the disparate parameters regardless of the units of measure for X or the magnitude of the numbers produced in those units.

Though quality index was devised to quantify the quality of varying batches of parts, Southwest Research Institute[1] recognized that measuring the fixed length of the many parts in a batch one time was not very different from measuring the varying temperature of a single oil multiple times. Traditionally, test developers have specified that certain controlled parameters (temperatures, pressures, flow rates, etc.) be at a certain target value plus or minus some tolerance. For the already stated reasons, it's not always very informative to compare the average of all data recorded to the limits of the tolerance. To get a more informative picture of the state of the control over the course of a test, Southwest Research a number of years ago began to calculate QI for each controlled parameter using the limits of the specification tolerance for U and L values.

[1] Southwest Research Institute is an independent not-for-profit research facility that, among many other things, conducts engine lubricant testing. It is located in San Antonio, Texas.

This approach works reasonably well so long as a few distinctions are kept in mind. First, the QI calculation was devised such that results range between zero and 1. Values less than zero do not occur because parts outside U and L have already been removed from the data by virtue of their failure to meet the specification tolerance. In collecting a stream of data from a running engine, however, there are no discrete parts to be selectively rejected; data above U and below L is allowed to occur with the resulting effect that the final QI number can be less than zero. Also, depending on the range of the tolerance, it may be necessary to calculate QI to six or more decimal places in order to distinguish between tests that examination of the plotted data would immediately show to be different. These factors combine to negate one of the key features of QI – the unitized scale. For example, using the specification tolerance for U and L, a value of 0.87 might be very good for fuel rate but completely unacceptable for coolant temperature. Recall the role of economics in determining tolerances and that many tolerances were determined in days when adjustments were made manually and infrequently. This is not to say that these specification tolerances are too wide, only that they are not ideally suited to use in QI calculation.

In order to keep all of the desirable features of QI intact, the constants used in the calculation must be more purposefully chosen. Consider these four examples:

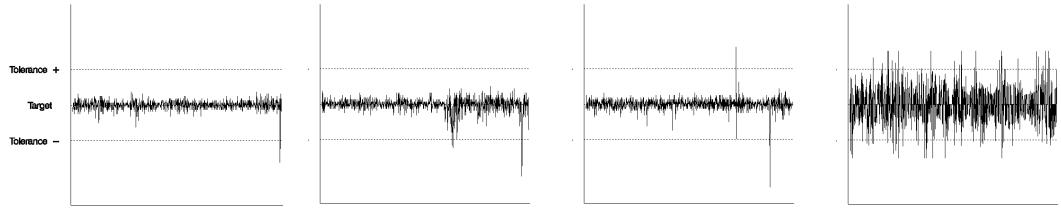


fig. h Plot 4 shows obviously different performance

If these plots are the output from four tests on the same stand, something was obviously different on the fourth run. In order for QI to be useful, it must indicate that. The fourth run may be different enough to be unacceptable. In order for QI to be useful, it must indicate that as well. And, it should do all this without needing data from the other three tests to provide context. So, how then to determine U and L constants that will produce calculated QI values that agree with the judgment arrived at by full review of the plotted data?

The ASTM Test Monitoring Center (TMC)[2] has developed a procedure that has been adopted for most of the new stationary engine tests of the last several years. Conceptually, the TMC approach is simple. The overarching goal for QI is that it be able to separate good tests from bad the same way that full review of plotted data would. So, the TMC approach is to review plotted data from a collection of representative tests and divide it into good and bad and then work backward to determine the QI calculation constants required to produce that result.

What tests are included in the "representative collection"? For a newly developed test, the data from the matrix[3] has typically been used. For reasons unimportant to this discussion, this data is usually contrived to be representative of what can be expected in future testing.

[2] The ASTM Test Monitoring Center is the organization that provides the test stand calibration system used by the lubricant testing industry. It is located in Pittsburgh, Pennsylvania.

[3] Matrix tests are those that are run as part of test development to assess the performance of the test while controlling for several key test variables (typically, different test labs, test stands, and oil formulations). Depending on the number of variables being considered, there are generally between 20 and 30 matrix tests generated for a new test.

The first step of the process is to plot the data for all of the controlled operational parameters for all of the tests and assess the performance of each one. Preliminary, pseudo-QI values for each test are computed using arbitrary U and L constants (usually the specification tolerance values as discussed previously in the Southwest Research approach). The tests can then be arrayed from the highest of these pseudo-QI values to the lowest. The general approach is to scan this array of plots and find the worst one that would still be acceptable. This will become the benchmark "zero test" that all future tests must match or exceed.

Properly selecting this zero test is the most critical part of properly constructing a QI system and requires appropriate engineering expertise. Often, the plot showing the lowest pseudo-QI really shows superior performance to the second lowest but for one or two or a handful of extreme data points. Plots showing these extreme data points are removed from consideration as the zero test because they are not representative.

On rarer occasions, there can be a handful of plots that show performance markedly different from all the others in the sample. If investigation shows a commonality among those plots (if they were produced by the same test stand, for example) and there is reason to believe that such performance will not be repeated in the future (if, in the course of generating those plots, the stand experienced a problem that has since been corrected, for example) then, likewise, none of those tests can be considered representative and can not be used for the zero test.

Most often, however, the data plots are continuously distributed from highest to lowest pseudo-QI. Where this is the case, the plot with the lowest pseudo-QI is used for the zero test.

The zero test is so named because in the next step, QI is set equal to zero for this test and its data is used to back-calculate the constants necessary to produce that result. Recall the equation for calculating QI:

$$QI = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{U + L - 2X_i}{U - L}\right)^2$$

Since the specification target for each controlled operational parameter is known and the constants are to be arranged symmetrically above and below it, U and L can be replaced as follows:

$$L = \overline{X} - \tfrac{1}{2}\Delta$$

$$U = \overline{X} + \tfrac{1}{2}\Delta$$

Where: $\overline{X}$ = specification target
Δ = difference between upper constant and lower constant

Substituting these values into the QI equation yields:

$$QI = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2\overline{X} - 2X_i}{\Delta}\right)^2$$

Setting QI equal to zero (which is the desired outcome for the zero test) and rearranging to solve for Δ gives:

$$\Delta = \sqrt{\frac{4}{n}\sum_{i=1}^{n}\left(\overline{X} - X_i\right)^2}$$

The summation data for the zero test is then used to solve for the required Δ. With Δ known, the QI for all other tests can be calculated using the above form of the QI equation and the results will all be relative to the benchmark zero test. When constructed following these steps, a QI system will have all of the desirable characteristics described earlier.

The foundation of this QI system is the collection of representative data upon which it is based. What happens if, as future testing proceeds, this collection turns out to have been not so representative? Recall that everything is based on the zero test which is the worst test that is still acceptable. In cases where the matrix collection is continuously distributed, there may well be future tests that are worse than the worst matrix test yet are still perfectly acceptable. These tests will compute a QI value less than zero and might be rejected.

For this reason, provision is made in the implementation of QI to allow sub-zero results for certain tests. Any time a sub-zero QI is calculated, the plotted data should be reviewed. If, as future testing proceeds, sub-zero QI tests come to be representative of what can be expected as normal, then the zero test must be reselected and the Δ for that parameter recalculated.

Other events can occur that will cause a sub-zero QI. It is possible for data to be recorded for one parameter that examination of other, interdependent parameters shows to be false. For example, an extreme coolant inlet temperature is likely to be erroneous if no corresponding anomaly is shown in the temperatures of the coolant outlet, inlet air, or exhaust. Loose wires and poor connections can sometimes cause recorded readings that are off scale or various electronics problems can cause data to be missing altogether. While any of these things can cause a sub-zero QI, these are not considered cases for recalculation of Δ because they are isolated incidents and not representative.

How can comparable QI values be calculated when data is missing or erroneous? It is possible for the same event to cause different data to be recorded on two different stands. A test stand in one laboratory might be configured to record a value of 9999 if a sensor fails while another lab might configure its stand to record, say, 100 for the same sensor failure. These two different numbers will have dramatically different effects on their respective QI results.

In cases like this, what is needed is agreement on what number will be reported when the actual value is unknown due to the acquisition system over- or under-ranging. These values are referred to as “floor” and “ceiling” values or under- and over-range values (or simply collectively as over-range values). They are so called because any values outside the over-range values are not accurately reported and are thus ignored.

How can data safely be ignored? Like the other constants used in QI calculation, the over-range values are carefully selected to make QI calculations give the desired result. The question that guides this selection is What values are so extreme that any value even more extreme is irrelevant in terms of its impact on the final QI value? The answer to this question is found by considering a perfect test. This test will have all of its data exactly at the targeted value and will have a final QI value of 1.000. Now consider a single one of the data points for this test moving off target. The further from target this data point gets the lower the QI for the test will become. How far from target can that single point in an otherwise perfect test get before the final QI value becomes negative? That point, on both the high and low side, is where the over- and

under-range values are set. There is no need to record values in excess of this since even a perfect test will have a sub-zero QI if as few as one of its points is greater (or less).

Computing the over-range values begins with the fact that the entire summation component of the perfect test with one extreme point is accounted for by that one point alone. This reduces the Δ form of the QI equation to:

$$QI = 1 - \frac{1}{n}\left(\frac{2\overline{X} - 2X}{\Delta}\right)^2$$

Because the QI in this case will be zero and Δ is known from previous calculation, this can be rearranged to solve for X, the extreme values:

$$X = \overline{X} \pm \tfrac{1}{2}\Delta\sqrt{n}$$

The two solutions to this equation will be the over- and under-range values. It is not important that these values make sense from an engineering standpoint. For example, it can occur that the under-range value for engine speed is negative. Obviously, an engine is not going to reverse direction. It is, nonetheless, important to resist the urge to "rationalize" these values to more intuitive ones. Doing so will alter the behavior of the QI system. Remember, the over-range values are only intended for use in data reporting and QI calculation and are completely independent from any system used for test stand control.

In cases where data is missing, how can the data be accurately assessed in comparison to a data set that is complete? The simplest approach would be to just calculate the QI with the data that is present. The problem with this approach, though, is that a test that records only 5% of its data will produce a QI as good (or bad) as that 5%. The remaining 95% of the test is completely unrepresented. This is, admittedly, an extreme case but it illustrates the point.

Some estimate of the performance of the test during the period of missing data is necessary. The best estimate of this period can probably be derived from the data gathered over the rest of the test. So, first the QI for the portion of the test having uninterrupted data collection is calculated using only the *n* points collected. This value is then scaled by the proportion of the test it represents (see the equation below).

$$QI_{missing} = QI\left(\frac{n}{n_{total}}\right)$$

Where: QI = QI for the data that is present
n = the number of data points present
n_{total} = the total number of data points that would be present in a complete data set

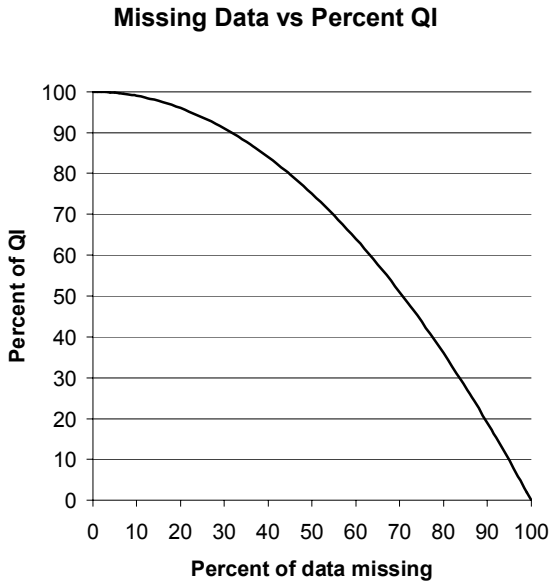
The portion of the test that is missing data is assumed to have performed at approximately this $QI_{missing}$ level. For the test as a whole, then, the $QI_{adjusted}$ is said to be the weighted average of the two portions of the test – the data-missing portion and the data-present portion.

$$QI_{adjusted} = \frac{QI * n + QI_{missing} * n_{missing}}{n_{total}}$$

Where: QI = QI for the data-present portion of the test

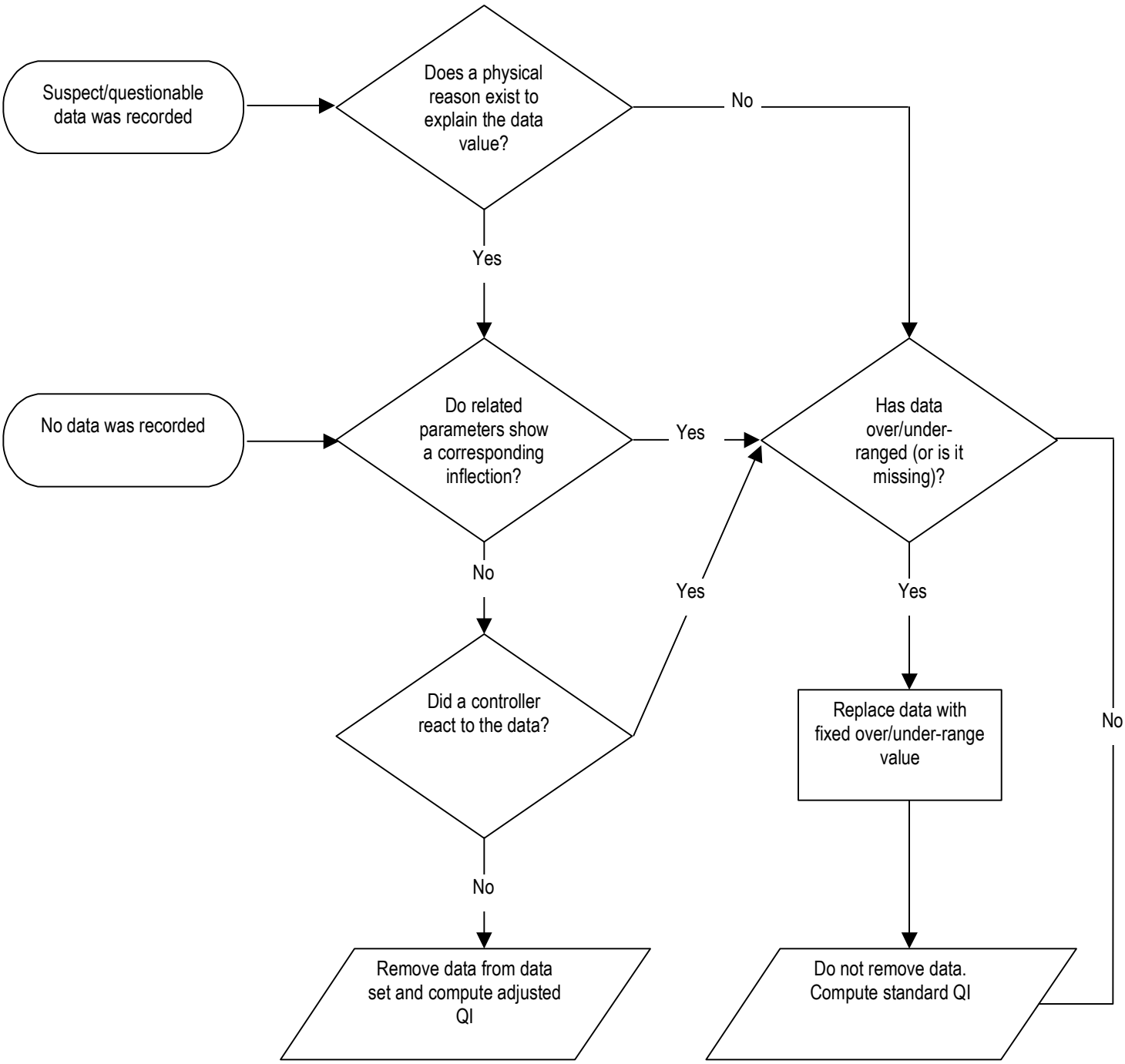
n = the number of data points present
$QI_{missing}$ = QI for the data-missing portion of the test
$n_{missing}$ = the number of data points missing
n_{total} = the total number of data points that would be present in a complete data set

The procedure for some test types places restrictions on the amount of data that can be missing. Caterpillar tests, for example, allow no more than four consecutive hours of missing data. The final report of DACA II[4] recommends that missing data be limited to no more than one percent of the total. The plot below shows the impact of missing data on the adjusted QI. For example, a test missing 30 percent of its data and having a QI of x for the remaining data would have an adjusted QI of 0.91x.



The flowchart below summarizes the steps for handling data that is either suspect or missing.

[4] Data Acquisition and Control Automation II was a subcommittee under section b of ASTM's committee D02 (on Petroleum Products and Lubricants) convened to make recommendations on automated data acquisition and control. It's final report was issued June 17, 1997.



Constructed as outlined here, Quality Index proves to be as useful in assessing collections of operational data as it is in gauging the physical characteristics of parts batches. Quality Index provides a complete, self-consistent way of quantifying the data presented in a plot. Of course, review of plotted data will always be important but the addition of QI can make assessment of operational control more uniform and objective. Since its introduction into the lubricant testing industry, QI has proven to be superior to any other statistical assessment of test control.