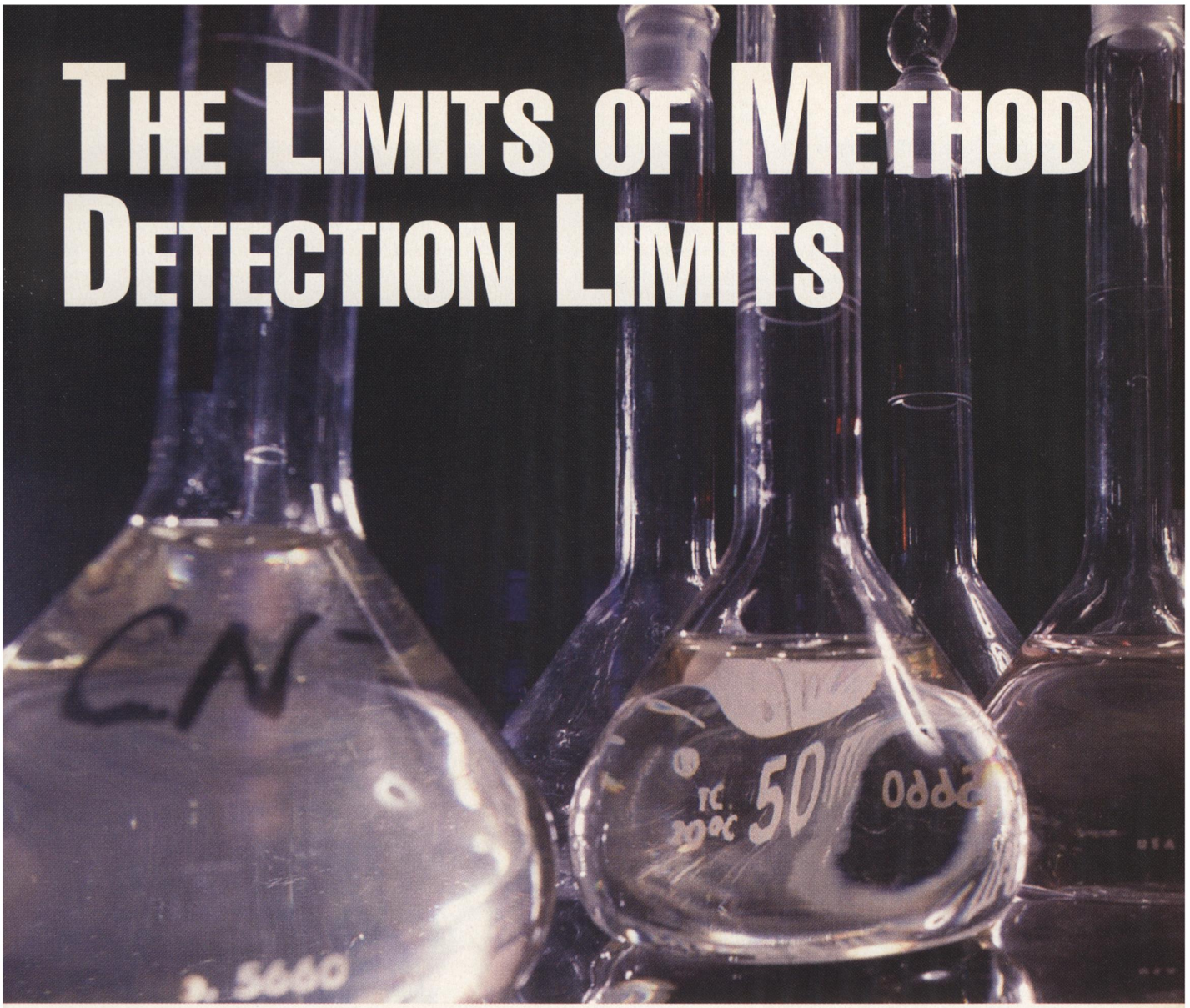
The Limits of Method Detection Limits



A simple, cost-effective alternative to the MDL procedure protects against false positives and other problems

Kenneth E. Osborn and Thomas Georgian

In March 2003, the U.S. Environmental Protection Agency (EPA) proposed revisions to the method detection limit (MDL) procedure in 40 *CFR* 136, Appendix B. Although the MDL procedure was established to determine the sensitivity of analytical test methods under the Clean Water Act, it has become the *de facto* industry standard for determining the detection capability of environmental test methods. Unfortunately, the proposed approach was seen as an extension of the previous approach and "remains a poor method to determine a laboratory's method sensitivity," according to the American Council of Independent Laboratories (Washington, D.C.).

Specifically, the existing and proposed MDL procedures have the following problems:

- The MDL underestimates long-term analytical variability. Since the MDL typically is determined from a single analytical event (a set of replicate samples processed in the same batch and analyzed the same day), it does not take into account the analytical variability that arises from different analysts, instrument calibrations, use of lots of reagents, and so forth.
- The MDL is not a conservative statistical limit that minimizes false positives. It protects against false positives only for the next single measurement, not for a large number of future measurements. The MDL is typically determined from only seven replicate measurements, giving rise to a statistical estimate of the detection limit that can vary from one determination to the next by a factor of about two.
- The MDL does not address false negatives. The proposed definition, which is not substantively different from the current one, states, "The MDL is an estimate of the measured concentration at which there is 99% confidence that a given analyte is present in a given

matrix." The MDL is essentially a "critical value" (a limit that minimizes false positives). If an analyte were absent in a sample, the probability that a single future sample measurement would be less than the MDL would be about 99%. Thus, if a measurement were greater than the MDL, the analyte would be reported as "detected" with at least 99% confidence. However, if a measurement were less than the MDL, the result could not be reported as "less than MDL" with a high level of confidence. If the analyte were present at a concentration near the MDL, the result would be erroneously reported as "less than MDL" (a false negative) about 50% of the time.

- The MDL does not take analytical bias into account, because it is calculated on the basis of only analytical precision. Neither upper nor lower acceptance limits are established for analytical bias. For example, it is primarily assumed that the mean concentration of a blank is zero. As a result, the procedure underestimates detection limits for test methods with a positive bias at low concentrations due to persistent blank contamination. The MDL could be substantially lower than the mean analyte concentration in method blanks.
- The MDL procedure can be expensive to perform because of the number of replicates potentially required to determine the MDL for each analytical method and instrument. For example, the proposed procedure states, "When developing an MDL for a new or revised method, or when developing a matrix-specific MDL, the MDL procedure must be iterated and the reasonableness of the MDL determined using an F-test." If the ratio of the variances for the two MDL determinations exceeds a specified value for the F-test, then the MDL procedure must be performed again until the critical value for the F-test is not exceeded.

The authors have devised an alternative procedure for calculating detection limits that overcomes most of the shortcomings of the current and proposed MDL procedures while remaining relatively cost-effective for routine environmental production work. Like the existing procedure, the approach uses a single-concentration-based design. While a calibration-based design would be expected to produce superior results, it would not be as cost-effective.

As in the approach used by the International Union of Pure and Applied Chemistry (Research Triangle Park, N.C.), two types of detection limits are used in the authors' approach:

- A "critical value," or detection limit that minimizes Type I error (false positives), denoted as L_C, and
- a detection limit that minimizes Type II error (false negatives), denoted as LD.

The LC detection limit is essentially a reporting limit for detections, and the LD detection limit is a reporting limit for nondetections. A measured concentration, X, would be reported as "detected" (at the 99% level of confidence) if $X > L_C$ and all method-specified identification criteria were met. Otherwise, the result would be reported as a "nondetect." LD defines the lowest reporting limit for nondetects. Thus, if $X < L_C$ or method-specific identification criteria are not met, the result would be reported as "$< L_D$."

Procedure

A detection limit study was performed as part of an initial demonstration of proficiency. At least seven low-level laboratory control samples (method blanks spiked at a concentration two to 10 times the estimated analytical detection limit) were processed through the entire test method. The seven replicates were analyzed in at least three separate analytical batches over 3 or more days to help take into account day-to-day sources of analytical variability. Replicate low-level laboratory control samples (LCSs) were used to calculate the sample standard deviation, *s*. For methods capable of reporting uncensored numerical results (for example, inductively coupled plasma spectroscopic trace metal methods), L_C was calculated using at least seven method blanks. However, because method blanks are routinely analyzed as batch quality control samples, a large number of method blanks (at least 30) is recommended to calculate *s*, since the ultimate objective is to estimate accurately the "true" standard deviation. Additional LCSs analyzed on a per-batch basis could also be used, as the sample statistic *s* becomes a better estimate of standard deviation as the number of replicates increases.

If fewer than 20 or 30 data points are available (the typical scenario), L_C is calculated as follows:

$$L_C = z_{1-p}\, s_{UCL,1-\gamma} = z_{1-p}\sqrt{[(n-1)/\chi^2_{n-1,\gamma}]}\; s \quad \text{(1a)}$$

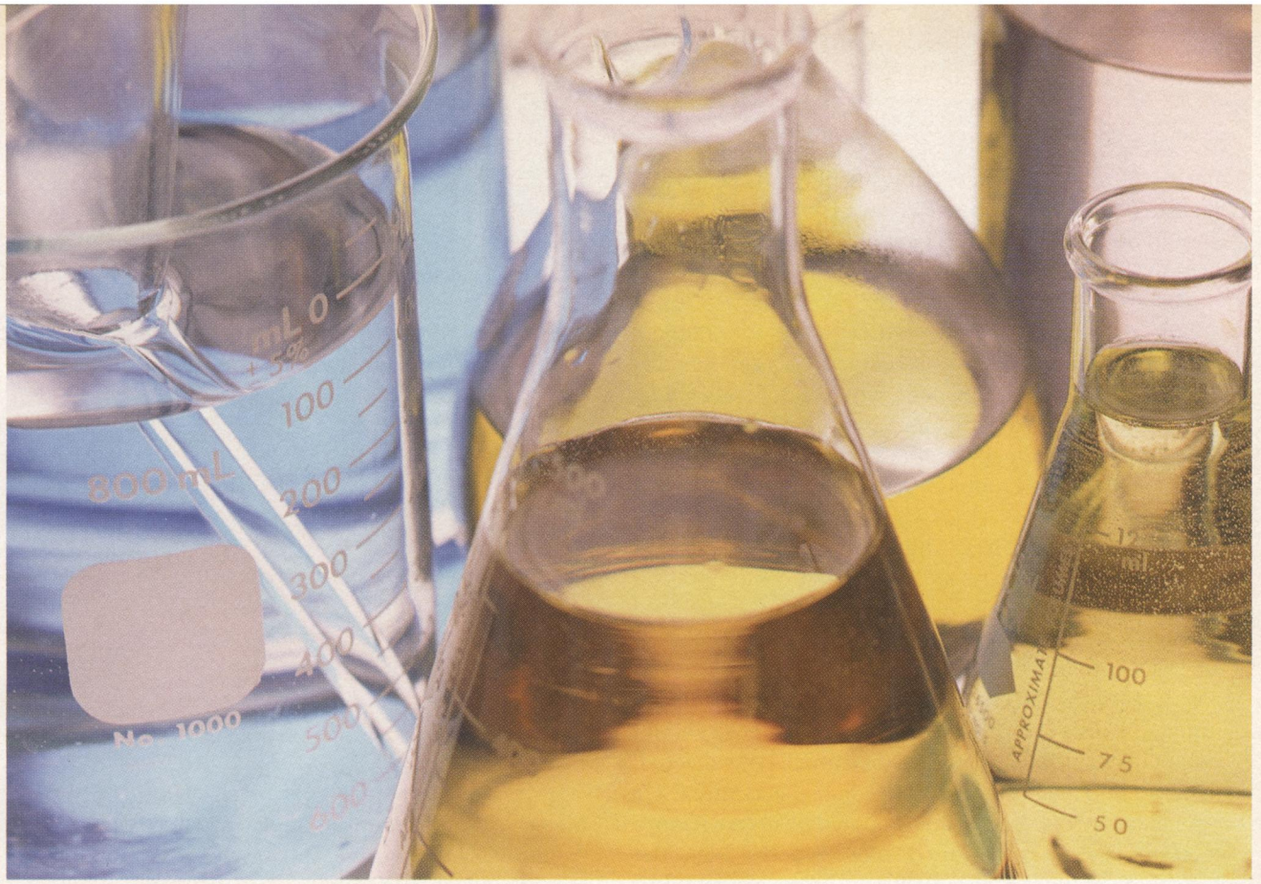
where

$\chi^2_{n} - 1,_{\gamma}$ is the $\gamma 100^{th}$ percentile of the χ^2 distribution, and

$z_1 - p$ denotes the $(1 - \gamma)100$th percentile of the standard normal distribution. (Note that, as in 40 *CFR* 136, normality is assumed.)

Equation 1a is the $(1 - \gamma)100\%$ tolerance interval that contains at least the proportion 1 - p of the population. If many analyses were performed using method blanks, then $(1 - p)100\%$ of the measurements would be less than L_C with $(1 - \gamma)$ 100% confidence. If $p = 0.01$ and $\gamma = 0.05$ (or 0.01), 99% percent of all future measurements would be less than L_C with 95% (or 99%) confidence.

$$L_C = 2.33\sqrt{(7-1)/0.872}\; s = 6.11\, s \approx 2MDL \quad \text{(1b)}$$



Equation 1b may be used in lieu of Equation 1a and provides a conservative estimate of L_C when $n > 7$. For a large number of replicates (more than 30), the formula in 40 *CFR* 136 may be used to calculate the L_C. However, it is recommended that Equation 1a be used even when $n > 30$, since the formula in 40 *CFR* 136 tends to underestimate L_C. It is also recommended that a statistical test for normality, as well as a test for outliers, be performed prior to calculating L_C.

False Negative Quality Control Sampling

L_C is verified using a "detection limit check sample" or "false negative quality control sample" (FNQS) rather than the iterative F-test procedure presented in 40 *CFR* 136. The FNQS not only verifies L_C but also establishes L_D — that is, becomes the lowest possible reporting limit for nondetects. The FNQS is prepared in the same manner as environmental samples. For example, for drinking water, the FNQS would be reagent water fortified with the analytes of concern at about two to three times the calculated L_C and then processed though the entire analytical method. However, when many analytes are being analyzed simultaneously, it may not be practical to prepare a spiking solution at two to three times the L_C for all analytes. Under these circumstances, the spiking concentration may be used for the L_D (FNQS) as long as it is two to 10 times L_C.

As stated earlier, the FNQS verifies L_C. Detection occurs if the measured concentration X is greater than L_C and all method-specific identification criteria are met. Once the FNQS is analyzed, it may be necessary to analyze additional FNQSs at higher or lower spiking concentrations. If an analyte in the FNQS is not detected, then L_C could have been underestimated, or a low bias could be present. The concentration of the FNQS should be increased until the analyte can be consistently detected (at least two consecutive FNQSs). If a significant negative bias were not present, then the L_C calculated from Equation 1a would be rejected, and L_C would be estimated using the FNQS spiking concentration. L_C would be set at half the concentration of the FNQS. However, if the nondetect for the FNQS were to arise from a large negative bias, then the calculated L_C would be retained, but L_D would be established from the lowest FNQS concentration that produces a detect.

If the response (signal-to-noise ratio) of the analyte in the FNQS is very high, then L_C may have been overestimated, and the FNQS concentration may be too high. The FNQS spiking concentration may be decreased to the smallest concentration that gives rise to detection. If two consecutive FNQSs produce detections, then the L_C may be reduced to half the lowest FNQS concentration that gives rise to detection. Reducing the FNQS to obtain a lower L_C or L_D value is optional. However, if analytical bias is not a significant factor and the FNQS is more than three times the calculated L_C, then half the concentration of the lowest FNQS (that gives rise to a detection) should be reported as an upper bound for L_C.

When the L_C calculated from Equation 1a is high relative to the value of L_D determined from the FNQS concentration, estimating $L_C = L_D/2$ should be done with caution. For example, if the calculated value of L_C (Equation 1a) is based upon a large number of values and L_D is based upon one or two FNQS analyses performed within a short

period of time, the FNQS analyses may not be a valid measure of L_D. Under these circumstances, FNQS data should be collected over a period of time before establishing L_D.

If the FNQS verifies L_C, then the reporting limit for L_D should be no less than the FNQS; that is, nondetects should be reported as "< Y," where "Y" denotes the FNQS concentration. However, when statistical evaluations of the data are performed, data censoring is undesirable. It is recommended that all numerical results be reported with the L_C and L_D values (a nondetect could be reported as "< Y [X]," where X denotes the measured value). One FNQS should be analyzed periodically (at least quarterly) as an on-going demonstration of L_C.

Optionally, as an additional check, a lab should analyze an FNQS per batch and calculate the recovery of the FNQS. After 20 FNQS results have been collected, the upper and lower control limits for the recoveries should be calculated. Until 20 FNQS results have been collected, initial control limits of ±50% are recommended. Any FNQS results outside of control limits should be noted. If more than 10% of FNQS results are outside of control limits, the L_C may have increased and should be recalculated. L_C has significantly increased (and should be revised) if the new L_C value is greater than twice the original.

Monitoring L_C Changes

Statistical tests that compare the degree of dispersion between two or more data sets could be used to monitor significant changes in L_C. For example, Levene's test could be used to compare the variance of a data set used to calculate an initial L_C value to the variance of a data set used to calculate a revised value. Levene's test is similar to Bartlett's test (F-test) for its homogeneity of variances but is more robust to departures from normality. Like Bartlett's test, if the Levene's test statistic exceeds a specified critical value, the baseline assumption that the variances of the two data sets are equal is rejected. If Levene's test indicated that the variances are significantly different, one could conclude that L_C has changed significantly. A new L_C value would be calculated and verified using an FNQS. Note that, regardless of whether a statistical test is used to monitor changes in L_C, an FNQS is essential in verifying L_C.

For analytical methods characterized by low-level blank contamination that cannot be completely eliminated, L_C is defined as the concentration that is statistically different from a method blank at the 99% level of confidence. Method blanks are used to calculate L_C when the mean blank concentration is significantly different from zero. At least seven method blanks (*not* spiked with any analytes) were analyzed. For n replicates, the L_C would be the 99% or 95% upper tolerance limit (γ = 0.99 or 0.95) for 99% coverage (p = 0.99):

$$L_C = \bar{x} + K_{\gamma,p,n}\, s \qquad (2)$$

where

$\bar{x}$ denotes the mean concentration for the method blank (note, once again, normality is assumed), and

K can be obtained from tables (for the noncentral t distribution) or estimated using the following formula:

$$K_{\gamma,p,n} \approx \frac{z_{1-p} + \sqrt{z_{1-p}^2 - ab}}{a} \qquad (3)$$

where

$$a = 1 - \frac{z_{1-\gamma}^2}{2(n-1)}$$

and

$$b = z_{1-p}^2 - \frac{z_{1-\gamma}^2}{n}$$

Note that when calculating L_C (especially using blanks), it is acceptable for some results to be negative. A simple conservative approach for establishing L_D would consist of multiplying the value of L_C calculated using Equation 2 by a factor of two. Alternatively, L_D could be established via the analysis of an FNQS, as discussed previously.

The proposed approach possesses a number of advantages over the existing MDL procedure. The approach protects better against false positives than the 40 *CFR* 136 MDL, since it more effectively takes into account long-term analytical variability and is based upon the tolerance interval for an unspecified number of future observations, rather than the prediction interval for the next observation. Furthermore, unlike the MDL, the approach addresses false negatives. The FNQS establishes the Type II detection level, the lowest reporting limit for nondetections, and provides an empirical verification of the Type I detection level. Unlike the MDL procedure, the Type I detection level is determined during method development only and is verified periodically via the analysis of an FNQS. Control charts can also be optionally generated for the FNQS recoveries to monitor method performance more effectively.

Kenneth E. Osborn *is a quality assurance officer at East Bay Municipal Utility District Laboratory (Oakland, Calif.) and* ***Thomas Georgian****, Ph.D., is a chemist with the U.S. Army Corps of Engineers in Omaha, Neb.*