Appendix B to the TSD
Characterizing Measurement Variability as a Function of Analyte Concentration for a Variety of Analytical Techniques

Summary

EPA's Office of Science and Technology (OST) is conducting a study of measurement variability as part of an assessment of detection and quantitation limit concepts. The study utilizes a large data set on measurement variability generated specifically to support this assessment and on data gathered during the course of analytical method development. The purpose of the study reported in this Appendix B to the Technical Support Document (TSD) is to determine if: (1) appropriate models could be found that would adequately characterize measurement variability and support the estimation of detection and quantitation limits; and (2) the number of models could be limited in order to minimize complications in the process of developing detection and quantitation limits. Results suggest that most measurement techniques used under the Clean Water Act (CWA) could be described by a general two-component model. The model describes measurement variability as a function of concentration in terms of two components: an additive component dominate at low concentration that represents variability to be independent of concentration and a multiplicative component dominate at higher concentrations that represents variability as proportional to the concentration of the substance measured. The model also allows for a smooth transition in the relationship in the concentration region where neither component is dominate. The study results also indicate that the two-component model may not be applicable to all analytes for all methods. In some cases, measurement variability does not exhibit a pattern that would suggest a particular model. In other cases, measurement variability is observed to be effectively constant over a range that extends beyond the low concentration range. In these cases, the two component model may still be a useful tool in determining detection and quantitation levels but other approaches, e.g., that do not involve the use of a particular model or assume that variation is effectively constant, may also be suitable.

1. Introduction

In 1998, EPA's Office of Science and Technology (OST) completed several studies conducted to generate data that would characterize measurement variability as a function of concentration for a variety of analytical techniques (see Appendix A). These studies were conducted as part of an effort to address criticism by the academic community, regulated industry, and others, of the detection and quantitation procedures used to support EPA's Clean Water Act (CWA) programs. More recently, the Agency has agreed to a schedule for completing this work and applying the results to the evaluation and selection of detection and quantitation concepts and procedures. This document uses the data referenced above and other selected data to characterize measurement variability as a function of analyte concentration for a variety of measurement techniques

This Appendix B to the Technical Support Document for the Assessment of Detection and Quantitation Concepts (TSD) was drafted before assignment of letters to appendices to the TSD. This resulted in duplication of appendix letters between the TSD and this appendix. To avoid confusion, references to the appendices in this Appendix B to the TSD are used without further defining words, whereas references to this appendix as an appendix to the TSD are used with the words "this Appendix B to the TSD." The appendices to this Appendix B to the TSD are contained on a CD-ROM that supports the TSD and are not physically attached to this document. A list of the appendices contained on the CD-ROM is provided at the end of this Appendix B to the TSD.

2. Data

EPA has conducted an extensive review of the published literature on detection and quantitation. Unfortunately, the published literature tends to focus on concepts and methodology and contains little actual data. No data were found on the key issue (for detection and quantitation) of the characterization and modeling of measurement variability as a function of analyte concentration. In order to fill the need for data, EPA embarked on an effort to produce data that would support the characterization of variability and the assessment of detection and quantitation. The first phase of this process was to develop data sets representative of the most commonly used analytical technologies, and the second phase was application of statistics, including model fitting, to these data.

Four data sets were selected to support the analyses presented in this Appendix B to the TSD. Three were developed by EPA for the express purpose of studying the relationship between measurement variation and analyte concentrations across a wide variety of measurement techniques and analytes. These data sets are referred to as (1) EPA's ICP/MS data, (2) EPA's Episode 6000 data, and (3) EPA's Episode 6184 data. In all three data sets, replicate measurement results from each combination of analyte and measurement technique were produced by a single laboratory over a wide range and large number of spike concentrations. The fourth data set was developed by the American Automobile Manufacturer's Association (the "AAMA Dataset") for the purpose of estimating a specific quantitation value referred to as an "alternate minimum level" (AML) [see Gibbons et al. (1997)]. For development of the AAMA Dataset, replicate results were measured at a limited number of spike concentrations by multiple laboratories using EPA Method 245.2 (CVAA) for mercury and EPA Method 200.7 (ICP/AES) for twelve other metals. Details of these four data sets are given in the subsections that follow.

2.1 EPA's ICP/MS Data

In 1996, EPA contracted with Battelle Marine Sciences to generate data to support the assessment of measurement variability in EPA's draft Method 1638 for nine metals by inductively coupled plasma with mass spectroscopy (ICP/MS). The nine metals were silver, cadmium, copper, nickel, lead, antimony, selenium, thallium, and zinc. The equipment used in this study made triplicate readings of each aliquot of each sample and averaged the results. Such averaging is common for ICP/MS design and use.

In preparation for this study, the ICP/MS instrument was calibrated using one sample aliquot at concentrations of 100, 1,000, 5,000, 10,000, and 25,000 nanograms per liter (ng/L). Initially, the calibration was performed using the default instrument software to produce unweighted least squares estimates assuming a linear calibration function. Subsequently, the analytical results were adjusted using weighted least squares estimates in place of the unweighted least squares estimates. Weighted least squares estimates are used to account for the fact that measurement variability may change as concentration changes. Usually, variability increases as analyte concentration increases over the full analytical range. Because the individual readings on each triplicate were not retained; the only data available to EPA are average intensity results for the aliquot. Draft EPA Method 1638 specifies the use of an average response factor rather than least squares estimation of linear calibration curves, although it does allow for the use of such procedures.

All nine metals were spiked into water to produce solutions at concentrations of: 10, 20, 50, 100, 200, 500, 1,000, 2,000, 5,000, 10,000, and 25,000 ng/L. Each solution was subsequently divided into 7 aliquots. The 7 replicates were measured in order of decreasing concentration. Seven replicates at 0 concentration (what chemists term a "blank") were also measured in triplicate for all nine metals. The resulting concentrations were converted to parts per trillion (ppt.) for ease of understanding and discussion.

Multiple instrument readings (at multiple mass to charge ratios or m/z's) were reported for most metals, although draft Method 1638 specifies only one acceptable m/z for eight of the nine metals. For lead, m/z's 206, 207, and 208 were all considered to be acceptable by draft Method 1638. The responses at the other m/z values for a particular metal are used for analyte identification purposes. This study of measurement variation used only data associated with m/z's that are specified by draft Method 1638.

2.2 EPA's Episode 6000 Data

Episode 6000 data were collected to characterize the variability of measurement results from 0.1 times an initial estimate of the Method Detection Limit (MDL) [Glaser et al (1981)] to spike concentrations 100 times the MDL. Measurement methods studied were:

- Total suspended solids (TSS) by gravimetry
- Metals by graphite furnace atomic absorption spectroscopy (GFAA)
- Metals by inductively-coupled plasma atomic emission spectrometry (ICP/AES)
- Hardness by ethylene diamine tetraacetic acid (EDTA) titration
- Phosphorus by colorimetry
- Ammonia by ion-selective electrode
- Volatile organic compounds in water by purge and trap capillary column gas chromatography with photoionization detector (GC/PID) and electrolytic conductivity detector (GC/ELCD) in series
- Volatile organic compounds by gas chromatography with a mass spectrometer (GC/MS)
- Available cyanide by flow-injection/ligand exchange/amperometric detection
- Metals by inductively-coupled plasma spectrometry with a mass spectrometer (ICP/MS)

Details of the study design are described in EPA's *Study Plan for Characterizing Variability as a Function of Concentration for a Variety of Analytical Techniques* (July 1998). The design is summarized below.

A method detection limit (MDL) study was conducted for each combination of analyte and analytical technique as an initial step in the generation of the Episode 6000 data. The study plan required laboratories to calculate these initial MDLs using the procedure in Appendix A. Seven replicates were then run at 100, 50, 20, 10, 7.5, 5.0, 3.5, 2.0, 1.5, 1.0, 0.75, 0.50, 0.35, 0.20, 0.15, and 0.10 times the initial MDL.

The following iterative procedure was used for organic compounds. Methods for organics normally list many (15 to 100) analytes, and the response for each analyte is different. Therefore, to determine an MDL for each analyte, the concentration of the spike must be inversely proportional to the response. The process of making spiking solutions with 15 to 100 different concentrations is complicated and prone to error. A more straightforward approach, and one used in this study, was to run 7 replicates at decreasing concentrations until signal extinction or until 0.1 times the initial MDL was reached, whichever came first, then select concentrations appropriate for the MDL.

Spike concentrations were measured in order from the highest concentration to the lowest to (1) to minimize carry-over effects and (2) to prevent the collection of data if when the instrument returned zeros for three or more of the replicates at a given concentration. Carry-over can occur when analysis of a high concentration sample is followed by a low concentration sample. Carry-over is usually less than one percent but can be a few percent in some methods. For example, if a sample at 100 times the MDL is followed by a sample at 0.1 times the MDL, the sample at 0.1 times the MDL could be compromised by carry-over because a small amount of carry-over from the 100 MDL sample could inflate the result for the 0.1 MDL sample. Running the samples in successive decreasing order should not affect successively lower measurements because the amount of carryover should be small relative to the measurement at the next lowest level.

Further details are described in EPA's *Study Plan for Characterizing Variability as a Function of Concentration for a Variety of Analytical Techniques* (July 1998).

2.3 *EPA's Episode 6184 Data*:

Episode 6184 data were generated to determine the concentration at which an analyte could no longer be identified as the concentration decreased. Details of the design for this study are described in EPA's *Study Plan for Characterizing Error as a Function of Concentration for Determination of Semivolatiles by Gas Chromatography/Mass Spectrometry* (December 1998). Data were generated for 82 semivolatile organic compounds by EPA Method 1625C (semivolatile organic compounds by GC/MS). MDLs were not determined for these compounds. Instead, solutions of the analytes were prepared at concentrations of 50.0, 20.0, 10.0, 7.50, 5.00, 3.50, 2.00, 1.50, 1.00, 0.75, 0.50, 0.35, 0.20, 0.15, 0.10, 0.075 and 0.050 ng/µL (or µg/mL). The solution at each concentration was injected into the GC/MS in triplicate with the mass spectrometer threshold set to 0, and again in triplicate with the mass spectrometer threshold set to a low level in the signal domain typical of that used in routine environmental analyses. As with the Episode 6000 dataset, samples were analyzed in order from the highest to the lowest concentration.

2.4 *AAMA Metals Data for EPA Methods 200.7 and 200.9 Data*:

The American Automobile Manufacturer's Association (AAMA) conducted a interlaboratory study of EPA Method 200.7 (metals by ICP/AES) and Method 245.2 (mercury by CVAA). Nine laboratories participated in the study, and each reported data for the following 13 metals: aluminum, arsenic, cadmium, chromium, copper, lead, manganese, mecury, molybdenum, nickel, selenium, silver and zinc. Study samples were analyzed by EPA Method 200.7 for 12 of the metals; mercury was determined by EPA Method 245.2.

The nine laboratories were randomized prior to the start of the study. Five matrix types (including reagent water) were selected, including four that were representative of the automotive industry. Each matrix was spiked at five concentrations in a predetermined concentration range. Matrix A (reagent water) was analyzed in all nine laboratories, and three laboratories analyzed each of the other four matrices. All analyses were repeated weekly over a five week period. As a result, a total of 6825 observations were obtained, which includes 2925 observations for matrix A (9 labs * 13 metals * 5 spike concentrations * 5 weeks) and 975 (3 labs *13 metals * 5 spike concentrations * 5 weeks) for each of the other four matrices (6825 = 2925 + [975 * 4]). There were two missing values for chromium in matrix A from labs 1 and 9.

Starting from a blank or unspiked sample, all target analytes were spiked at 4 concentrations to yield a total of five concentrations per matrix. Concentrations ranged from 0.01 to 10 ug/L for mercury and selenium, respectively, on the low end, and from 2.0 to 1000 ug/L for mercury and selenium, respectively, on the high end. In addition, the concentrations were matrix-dependent. The same concentration ranges for each metal by matrix combination were used for all five weeks of the study. A copy of the original study proposal is presented in Appendix C.

3. Statistical methodology

The study data were first analyzed graphically and then numerically. Numerical analyses consisted of model parameter estimation and evaluation.

3.1 Graphical Analysis Techniques

Composite plots for all combinations of analyte and analytical technique were produced. These sets include: (1) Measurement Results vs Spike Concentration (Appendix B.1), (2) Log_{10} [Measurement Results] vs Log_{10} [Spike Concentration] (Appendix B.2), (3) Observed Standard Deviation vs Spike Concentration (Appendix B.3), (4) Log_{10} [Standard Deviation] vs Spike Concentration (Appendix B.4),(5) Relative Standard Deviation vs Log_{10} [Spike Concentration] (Appendix B.5), and (6) Standardized Residuals vs Log_{10} [Spike Concentration] (Appendix B.6). These graphics are contained on the CD-ROM in the file Plot_EPA_AAMA.pdf.

The purpose of these plots is to allow examination of the relationship between two numerical variables and determine if the data fall close to a curve describing an expected model.

(1) The first set of plots of measurement results versus spike concentrations can be used to evaluate the mean recovery model. If the assumed linear model were true then the relationship outlined by the plotted data would be approximately linear.

(2) A plot of log measurement results versus log spike concentration will show an approximately linear relationship, similar to (1), if the relationship between measurement results and spike concentrations is well behaved. Otherwise, the log transformations used in these plots will tend to exaggerate deviations from the linear model. In particular, a positive bias in measurement results will show as an almost flat line at low concentrations before it starts to increase as concentration increases. A negative bias will show as an almost perpendicular drop at the lower concentrations. The primary advantage of the log-log plot is that it allows for easier visualization of the relationship between individual measurement results spike concentration at the low concentrations. When the log-log plot is combined with an indicator that identifies which data came from which calibration, the effect of changing calibration may be seen clearly.

(3) The plot of observed standard deviations versus spike concentrations can be used to evaluate the reasonableness of the constant and/or straight line models. If the constant model for standard deviation were true then the standard deviation would be approximately the same regardless of concentration. If the straight line model for standard deviation were true then the plots are expected to indicate an approximately linear relationship.

(4) The log-log plot is expected to display an approximately linear relationship when a logarithmic model fits the data. If the two-component model were true then we would generally expect to see a relationship that looks something like the shape of a hockey stick. That is, standard deviation would be approximately constant at low concentrations and would increase in proportion to concentration at higher concentrations. However, if spike concentrations were not selected at sufficiently low concentrations, the two-component model would display the approximately linear relationship of a log-log model.

(5) The plots of relative standard deviation (RSD) versus log 10 of the spike concentration are generally expected to show high RSD for the lowest concentrations and convergence to a constant RSD at higher concentrations. This theory may be used to describe measurement variation in terms of a single number.

(6) The plots of standardized residuals versus log 10 of the spike concentration show how well variation is estimated using the available procedures for the two component model. A residual shows how much a measurement result has deviated from the model of the relationship between measurement results and spike concentrations. Standardized residuals are generally expected to be within plus or minus 3 standard deviations of zero and are expected to average zero. Residuals at each spike concentration are standardized to the standard deviation estimated using the two-component model. When residuals spread to less than three standard deviations, the model is generally overestimating variability. When residuals spread to more than three standard deviations, the model is generally underestimating variability. The exception to this rule is an apparent outlier. A secondary characteristic of this type of plot is that it accentuates any deviation from the expected linear relationship.

3.2 Model Estimation

With some deviations, the two component model (Rocke and Lorenzato,1995) was estimated using the FORTRAN program provided (publicly available at no cost) by Professor David Rocke (http://handel.cipic.ucdavis.edu/~dmrocke/). Qualitative output and plots of deviations from this model were used to evaluate the fit of the model to the available data. The model is expressed as:

$$y = \alpha + \beta\mu e^{\eta} + \varepsilon$$

where the parameters are defined as:

y	= observed instrument response
e	= the natural exponential function
α	= the intercept of a linear function
β	= the slope of a linear function
ε	= independent errors with normal distribution, mean zero, and fixed variance
η	= independent errors with normal distribution, mean zero, and fixed variance
μ	= true concentration in the sample measured

In this Appendix B to the TSD, two deviations from the two-component model are present. These deviations are that results have been substituted for instrument responses and all data used in this Appendix B to the TSD have entrained dependencies that violate the assumption of independent errors within the domain of interest (e.g., within a single laboratory or within a group of laboratories). For analytes and analytical techniques for which results are generated by a linear response, substituting results for responses does not fundamentally change the structure of the model. With regard to independent errors, results for each combination of analyte and measurement technique were obtained in descending order of concentration. Hence, any event that happens in the measurement process during data collection has effects that systematically continue throughout the remaining portion of data collection. The most obvious event is calibration, which can be linked visually to systematic changes in results. It is not clear whether the AAMA data have the same problem. However, the AAMA data were collected in sets spaced one week apart and each set contained all of the spike concentrations used for every analyte under study. Hence, any event that affects any one result at a given spike concentration is likely to affect other results at every other spike concentration. Again, the obvious event is calibration.

Model estimates were restricted to cases for which at least 5 spike concentrations were available. Because the two-component model has 4 parameters, the extra spike concentration is required to have a real indication that the model is fitting any given dataset.

4. Results

4.1 Graphical Evaluations

In this section, we provide some examples of diagnostic plots. Graphics referred to in the text are shown in Section 4.1.7. Complete sets are provided in graphics files on the compact disk that contains the attachments to this document.

4.1.1 Results vs. Concentration

In general, the observed relationship between concentration and results follows a straight line closely (see, e.g., Figure 4-1 which shows measurement results versus concentrations for WAD Cyanide). In other cases, these plots help identify analytes with potential outlier observations (see, e.g., Figure 4-2 which shows measurement results versus concentration for 1-Chlorobutane). In still others, the linear relationship is not well demonstrated in the EPA data (see, e.g., Figures 4-3 and 4-4). Although more difficult to visualize because of fewer spike concentrations, much of the observed AAMA data also appear to follow a straight line (see, e.g., Figure 4-5 Cadmium data from lab 3). Because one complete set of spike concentrations were measured each week, some of the AAMA data show an effect related to the week the measurements were made. An extreme example is shown in Figure 4-6 Copper from lab 4. The highest copper measurement result at each spike concentration was made during week 2 of the study. The complete set of plots for measurement results vs. spike concentrations is provided in Sections B-1-1 and B-1-2 of the file Plot_EPA_AAMA.pdf contained on the compact disk.

The number of low concentrations included in the EPA data made possible the examination of patterns of variation in the low concentration range. The rationale for this was the importance of variation in the low concentration range from theoretical and practical perspectives. We specifically focused on the lowest six concentrations reported in EPA studies. The selection of six was based on the judgment that six was large enough to determine a departure from the linear model and small enough to visually resolve the individual results given the range of the concentration levels. In general, the plots show patterns of variation not related to concentration. This may be considered evidence of the constant variation property in the low concentration and also may be due to the fact that the low concentration levels in the EPA studies were designed to demonstrate the performance of methods below their established detection levels. Examples of these plots are shown in Figures 4-7 and 4-8, antimony and 2,6-dinitrotoluene.

4.1.2 Log_{10} [Measurement Result] vs Log_{10} [Spike Concentration]

For the EPA data, these plots indicate that discontinuities in the relationship between results and spike concentrations can occur when calibration is known to change (see, e.g., Figure 4-9, plot for 1,3-dichlorobenzene and Figure 4-10, 2-chlorotoluene). No remarkable relationships were noted among the AAMA data, though it would be difficult to discern any pattern with only five spike concentrations. The complete set of plots for Log_{10} [Measurement Results] vs Log_{10} [Spike Concentration] can be found in Sections B-2-1 and B-2-2 of the file Plot_EPA_AAMA.pdf contained on the compact disk.

4.1.3 Observed standard deviation vs. spike concentration

These plots generally indicate that measurement variation at low concentrations may be approximately constant (see, e.g., Figure 4-11, sodium). However, it is unusual for measurement variation to remain constant throughout the range of spike concentrations considered in the EPA and AAMA studies. At some point, there is generally an indication that measurement variation increases with spike concentration. In certain cases, the plots provide some indication that the relationship between variability and concentration may be increasing linearly (see, e.g., Figure 4-12, Silver [Ag] 107). On the other hand, other graphics indicate both convex and concave curves ((Figures 4-13, 2-chorophenol and Figure 4-14, 1,4-dicholobenzene). With only five spike concentrations, the AAMA data do not appear to be informative in these plots. The complete set of plots is provided in Sections B-3-1 and B-3-2 of the file Plot_EPA_AAMA.pdf.

4.1.4 Log/log plot (standard deviation vs spike concentration)

The log-log plots of standard deviation versus spike concentration display three general patterns. In the first, measurement variation appears random within an approximately oval shaped regions that does not suggest a relationship between log standard deviation and log concentration (Figure 4-15 WAD Cyanide). In the second, lower concentrations exhibit roughly constant variability within an oval shaped region but higher concentrations show an increasing linear relationship between log standard deviation and log concentration (Figure 4-16 Ammonia as Nitrogen). In the third, all concentrations exhibit log standard deviation that increases linearly with the log concentration (Figure 4-17 Mercury). Note that the approximately linear, concave, and convex relationships shown in the figures cited in Section 4.1.3, above, all display as approximately linear in the log/log plots (Figure 4-18 Silver [Ag] 107, Figure 4-19 2-chorophenol and Figure 4-20 1,4-dicholobenzene). This suggests that the apparent relationships shown in the standard deviation versus concentration plots were artifacts of a log/log relationship that is approximately linear. The complete set of plots is provided in Sections B-4-1 and B-4-2.

4.1.5 Relative Standard Deviation versus Spike Concentration

Relative standard deviation (RSD) generally appears to decrease as concentration increases throughout the range of concentrations observed in this study although there are some exceptions (e.g., Figure 4-21 Chlorobenzene and Figure 4-22 Dibromoethane). The more typical pattern is shown in Figure 4-23 1-Chlorobutane. RSD may approach asymptotically some limiting value but it does not generally appear to become constant at any point. Note that the range on the vertical axis changes considerably between plots and that the larger ranges will tend to hide the variation that continues to exist at higher concentrations. For depicting method performance, it may be appropriate to select the lowest acceptable result, calculate the RSD, and describe the method as being capable of meeting or doing better than the selected standard. Cases where RSD increase with increasing concentration may be problematic.

4.1.6 Summary of the Graphical Analyses

The primary purpose of graphing the study data was to support the selection of models for relating the variability of results to spike concentrations. However, artifacts of the way data were collected are also visible in these analyses.

The classes of models considered include: (a) constant variability (not related to spike concentration), (b) variability increasing linearly with concentration, (c) log variability increasing linearly with log concentration, and (d) variability increasing proportionally with concentration after shifting upward some fixed amount. The data seem to indicate that model class (d) is able to describe the relationship between results and concentrations across a wide variety of measurement techniques and analytes. One specific model in this class is the two-component model described by Rocke and Lorenzato (1975), among others. Model classes (a) and (c) can be considered subsets of model class (d) that appear when concentrations do not cover the full range of the method.

The effect of using systematic elements in the sampling design, as opposed to using completely random sampling, is present in all datasets considered here and clearly demonstrated in the Episode 6000 data and the AAMA data. In the Episode 6000 data, all results for each combination of analyte and measurement technique were measured in order from the highest to the lowest. Where significant changes in the measurement process occurred, such as instrument re-calibration, visually observable effects in the relationship between results and concentration became apparent. In the AAMA data, every spike concentration in the study was measured once a week for five weeks. Where significant changes in the measurement process occurred between weeks, visually observable effects in the relationship between results and spike concentration became apparent.

4.1.7 Graphics

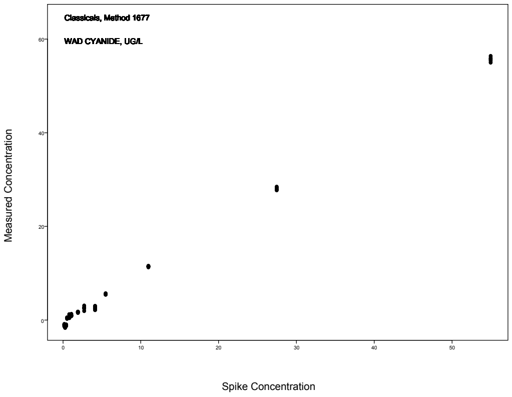

Figure 4-1

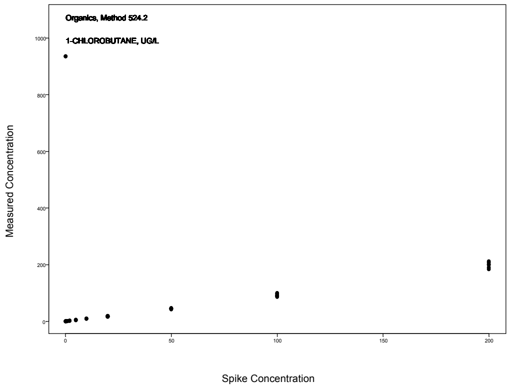

Figure 4-2

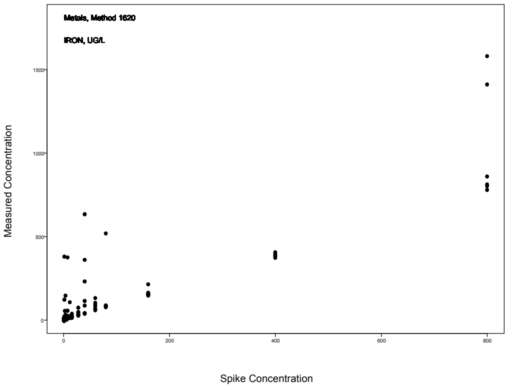

Figure 4-3

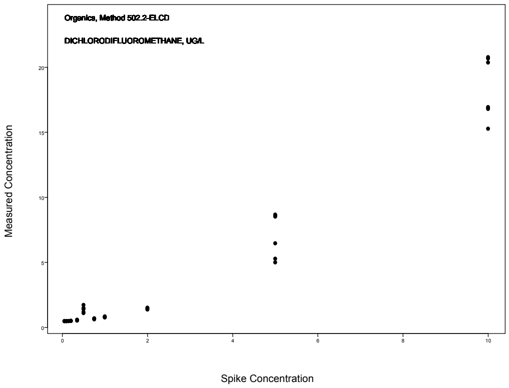

Figure 4-4

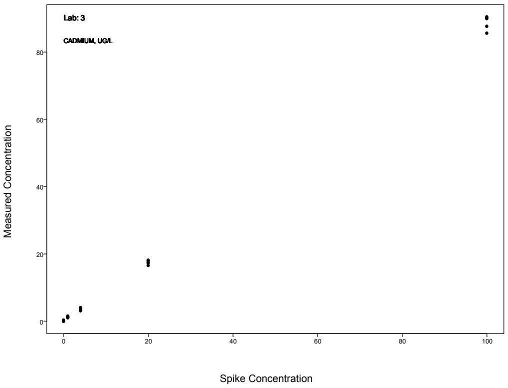

Figure 4-5

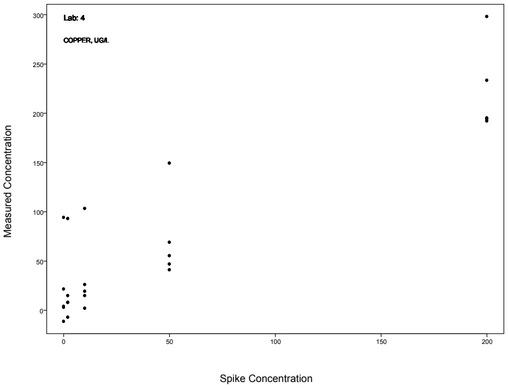

Figure 4-6

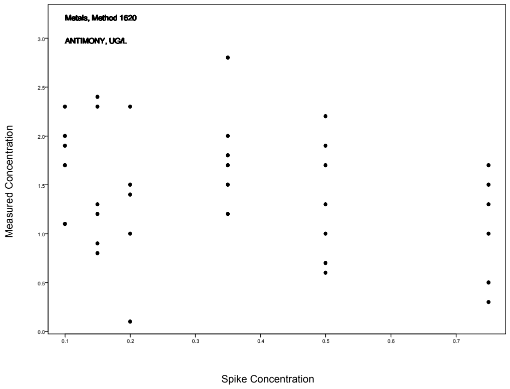

Figure 4-7

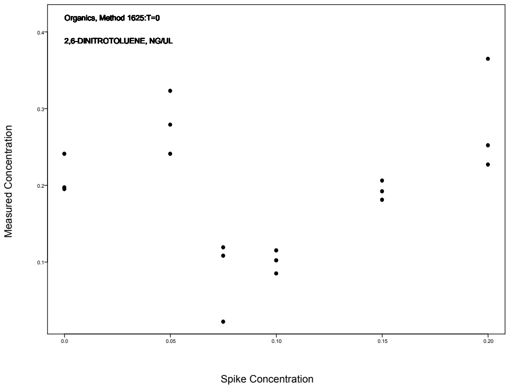

Figure 4-8

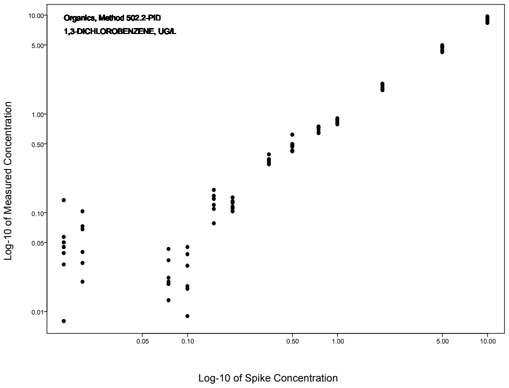

Figure 4-9

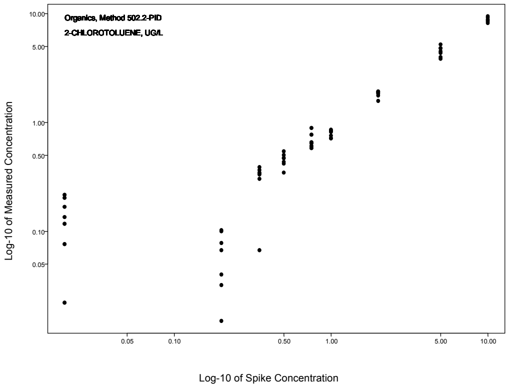

Figure 4-10

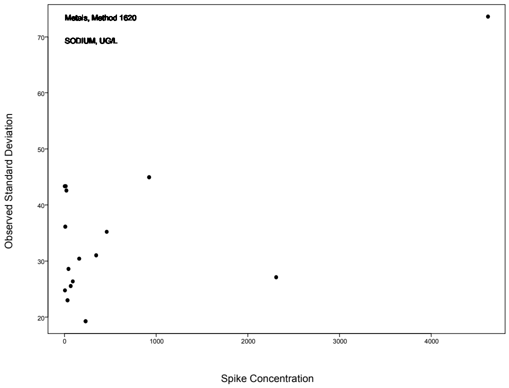

Figure 4-11

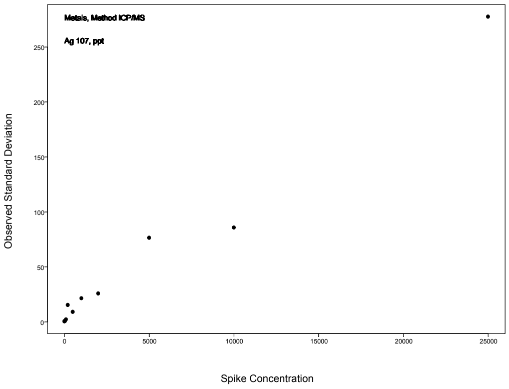

Figure 4-12

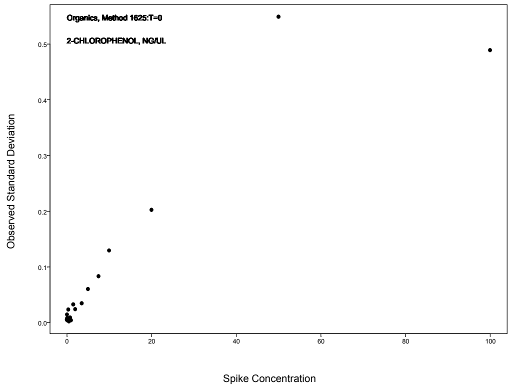

Figure 4-13

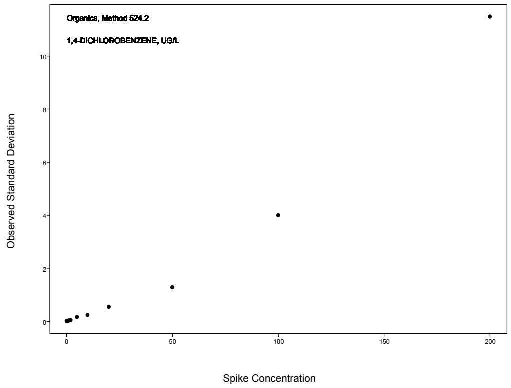

Figure 4-14

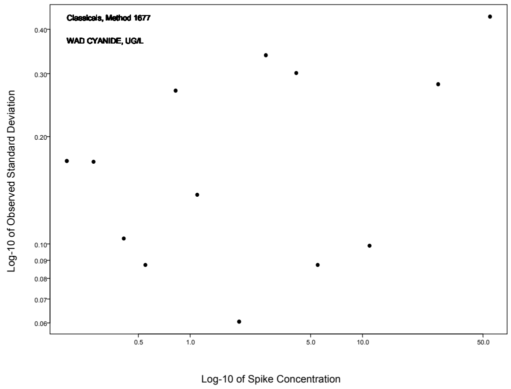

Figure 4-15

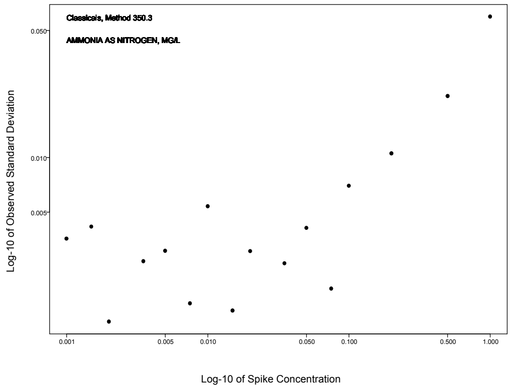

Figure 4-16

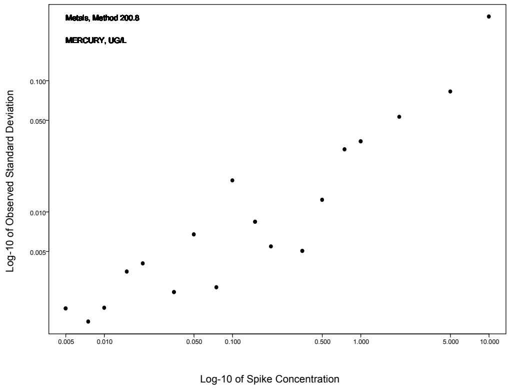

Figure 4-17

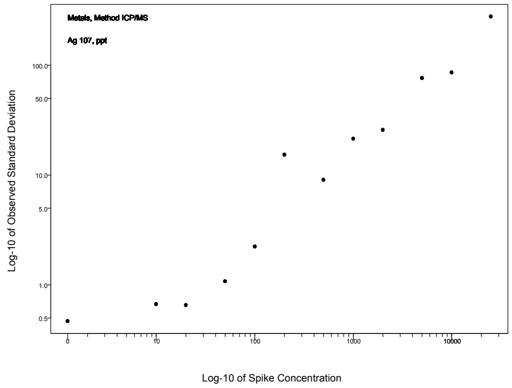

Figure 4-18

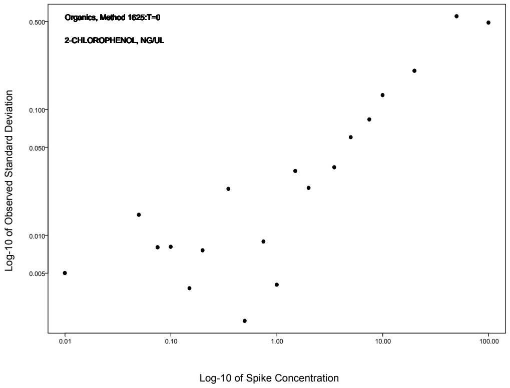

Figure 4-19

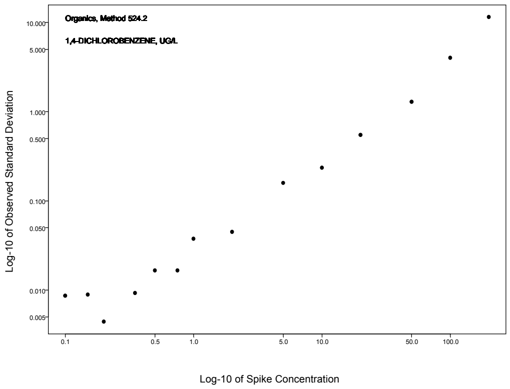

Figure 4-20

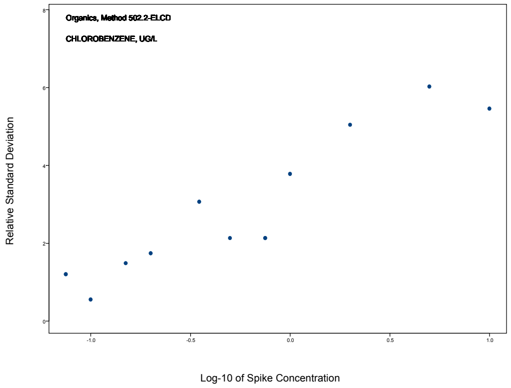

Figure 4-21

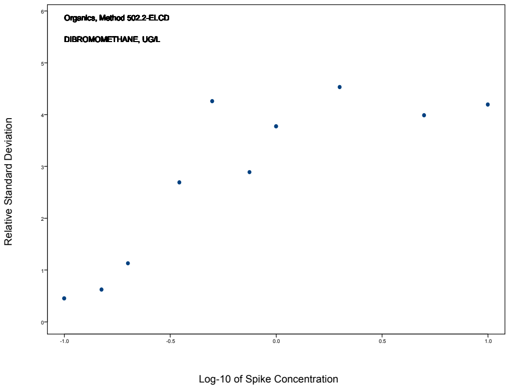

Figure 4-22

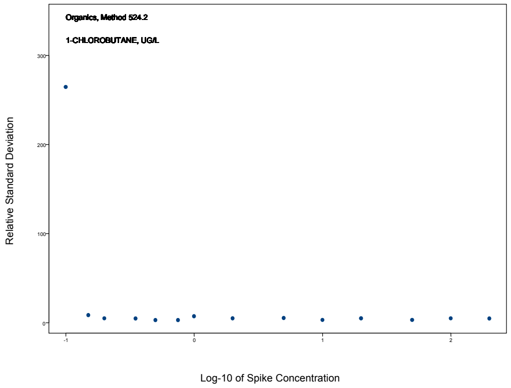

Figure 4-23

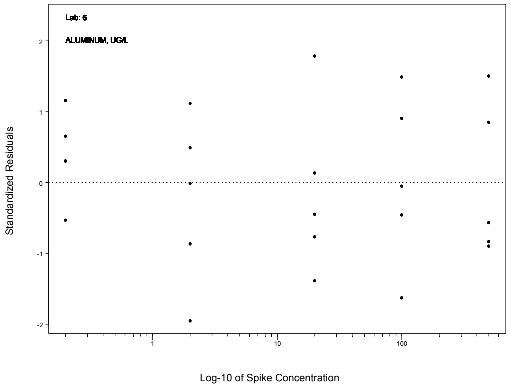

Figure 4-24

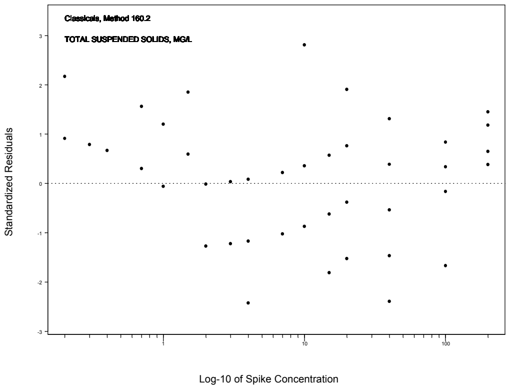

Figure 4-25

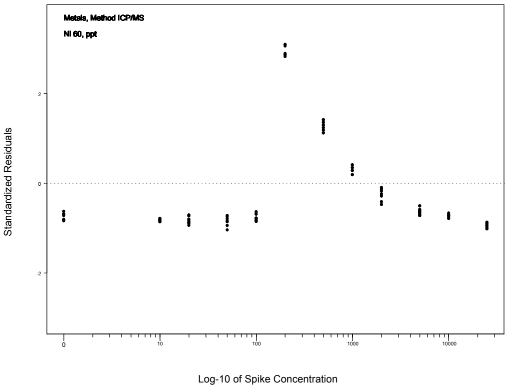

Figure 4-26

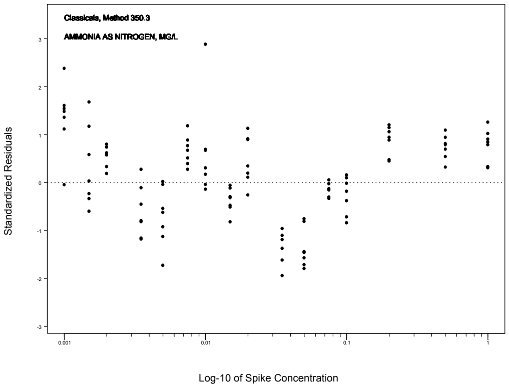

Figure 4-27

4.2 Two-component Model fitting

This section considers the results of fitting the two-component Rocke-Lorenzato Model to the EPA and AAMA. The basic approach was to use maximum likelihood estimation to estimate parameters of the Rocke-Lorenzato model using software developed by Professor Rocke which is available at no cost on his web site http://www.cipic.ucdavis.edu/~dmrocke/software.html. The software implements a numerical optimization algorithm that solves for the maximum likelihood estimates of the parameters of the model. The results for the EPA data are summarized in Table 4-1 which lists the number of analytes by method and the number of times the software was run and the number of times that the software was not able to obtain a solution, i.e., failures. Failure to obtain a solution is usually the result of poor fit of the data to the data. That is, the data are so discrepant with regard to the assumed model that the algorithm fails to converge to a solution. For the EPA data, 47 of 371 cases (about 13%) failed to obtain a solution. There were no failures for the AAMA data. The appendices to this document contained on the companion CD-ROM include tables with results of the maximum likelihood estimation including parameter estimates along with recorded optimization failures by study, class, technique, and analyte. Plots of residuals, useful for evaluating model fit are shown in Plot_EPA_AAMA.pdf which is also on the CD-ROM.

Table 4-1: Optimization Failures for Two-Component Model Estimates

EPA Study	Measurement Method	Number of Analytes Measured (Analytes x Labs)	Number of Failures
Episode 6000	130.2	1	0
Episode 6000	160.2	1	0
Episode 6000	1677	1	0
Episode 6000	350.3	1	0
Episode 6000	365.2	1	0
ICP/MS	ICP/MS	11	2
Episode 6000	1620	26	0
Episode 6000	200.8	21	9
Episode 6000	502.2	63	4
Episode 6000	524.2	80	0
Episode 6184	1625	167	32
Total		371	(13%) 47

The two-component model can be judged to fit the EPA and AAMA data since the maximum likelihood algorithm was able to obtain parameter estimates for the large majority of the data sets. Evaluation of model fit should also include examination of residual plots. The two-component model appears to fit the AAMA data well but the fit is inconsistent for the EPA data. For the AAMA data, the maximum likelihood algorithm had no difficulty obtaining parameter estimates for the two component model and plots generally indicate the expected random pattern of the residuals about zero (e.g., Figure 4-24 Aluminum, Lab 6). For the EPA data, some analytes have residual plots that indicate a reasonable fit and others have residuals that indicate questionable fits. For example, the results indicate that the fit for Method 160.2 seems reasonable (Figure 4-25 Total Suspended Solids). For other methods in the EPA studies, the model fit is inconsistent with some analytes showing reasonable residuals and others indicating problems with model fit (e.g.Figure 4-26 Nickel 60). For many of the analytes it was possible to assess the effect of changes in calibration on model fit. For a number of analytes, plots of results vs. concentration were generated by calibration sequence. For example, residuals shown in Figure 4-27 Ammonia as Nitrogen indicate the presence of four different calibrations. The complete set of plots is presented in Plot_EPA_AAMA.pdf. Also, in addition to performing diagnosis and analysis based on the full set of available spike concentrations for each analyte, model parameters were estimated based on calibration sequences. Tables contained in the appendices contained on the companion CD-ROM to this document show the estimated parameters by calibration sequence. Information was not available to allow diagnostics and the plotting of organics by calibration sequence.

5.0 On Designing Studies of Results versus Concentration

Based on evaluation of the data generated, four improvements are suggested to the original EPA study design.

The first suggestion is directed towards instruments that report no value or a constant value below some fixed spike concentration. In these cases, perform a preliminary assessment of instrument capabilities prior to selecting the lowest concentrations for the tests. The assessment would include a limited number of measurements at low concentrations to determine the capability of the instrument to generate low level measurements. Then, in conducting the study, restrict the selection of spike concentrations in the experiment to those concentrations high enough to generate suitable measurements.

A second suggestion would be to include replicate measurements at increasing concentrations until the upper end of the calibration range is achieved. Data at the highest possible concentrations would allow enhanced definition of the proportional error region.

The third suggestion would be to create all study samples and aliquots from the samples prior to measurement and to completely randomize the order in which aliquots are measured. The possibility of carryover would exist but it would be handled the same way it is handled by a laboratory using a measurement method in production, i..e, by re-analysis of a low concentration sample that follows analysis of a high concentration sample after analysis of a blank to demonstrate that the analytical system is essentially free from carryover. This could increase the analytical cost by an order of magnitude.

The fourth suggestion would be to use multiple laboratories. This assumes that inter-lab objectives are consistent with study goals. Of course, depending on how many and which measurement methods are selected for study, the cost of such a study using multiple laboratories could easily run into several million dollars.

6.0 Conclusions

The study data suggest that measurement variation for most measurement techniques used under the Clean Water Act can be described by two models. Those models are the two-component model and the constant model. Though the pure lognormal model could be used to fit data that do not show a constant component, this model implies the unacceptable condition that measurement variation may become zero at some low concentration. Graphics of the variability associated with EPA's data combined with both the graphics and the two-component estimates from the AAMA data makes the suggestion of these two models fairly strong for metals data. Although the two-component model has an appealing physical basis (see Rocke and Lorenzato [1995]), the available data do not provide strong support for or organic analytes.. Graphics of variability versus concentration in many cases suggest the two-component model but the actual fit of this model using the maximum likelihood algorithm sometimes fails. We assume that the cases where the two-component model does not work well for the organics data in this study are due to the lack of a response at low concentrations or because of problems in the design and implementation of the procedures used to generate the data.

Unfortunately, problems associated with the design of the studies considered in this assessment make it unclear whether the estimation procedures associated with the two-component model can be relied on to routinely produce estimates under the necessary conditions. More than 10% of the analyte/measurement technique combinations considered here have failed to produce maximum likelihood estimates for the two-component model. More of these combinations appear to fail the graphical examination of the fit. For our purposes, the true test of the model is how well it can be used in practice to produce detection or quantitation estimates. If a simpler model produces estimates on a more reliable basis, trade offs in terms of precision and bias would have to be considered.

References

(1) **EPA** (1995), Appendix B to Part 136 - Definition and procedure for the Determination of the Method Detection Limit - Revision 1.11, *40 Combined Federal Register*, U.S. Goverment Printing Office, Washington.

(2) **Glaser, J.A., Foerst, D.L., McKee, G.D., Quane, S.A., and Budde, W.L.** (1981), Trace Analyses for Wastewaters. *Environmental Science and Technology*, **15**, pp.1426-1435.

(3) **Rocke, D.M., and Lorenzato, S.** (1995), A Two-Component Model for Measuremnt Eroor in Analytical Chemistry, *Technometrics*, **69**, pp.3069-3075.

(4) **Gibbons, R.L, Coleman, D.E. and Maddalone, R.F.** (1997), An Alterntive Minimum Level Definition for Analytical Determination, Environmental Science and Technology, **31**, pp.2071-2077.

Appendix B Appendices and Tables on CD-ROM

Graphics (Sorted by Study, Class, Measurement Technique) in Plot_EPA_AAMA.pdf

- B.1. Measurement vs. Spike Concentration
  - B.1.1 EPA
  - B.1.2 AAMA
- B.2 Log Measurement vs. Log Spike Concentration
  - B.2.1 EPA
  - B.2.2 AAMA
- B.3 Observed Standard Deviation vs. Spike Concentration
  - B.3.1 EPA
  - B.3.2 AAMA
- B.4 Log Standard Deviation vs. Spike Concentration
  - B.4.1 EPA
  - B.4.2 AAMA
- B.5 Relative Standard Deviation (RSD) vs. Log Spike Concentration
  - B.5.1 EPA
  - B.5.2 AAMA
- B.6 Standardized Residuals vs. Log Spike Concentration
  - B.6.1 EPA
  - B.6.2 AAMA
  - B.6.3 EPA (by Calibration)

Tables (Sorted by Study, Class, Measurement Technique) in wpd files

- Table 1_AAMA.wpd
- Table 1_EPA.wpd
- Table 2_EPA.wpd