United States Environmental Protection Agency | Office of Water (4303) | December 2011

A Laboratory Study of Procedures Evaluated by the Federal Advisory Committee on Detection and Quantitation Approaches and Uses in Clean Water Act Programs

December 2011

U.S. Environmental Protection Agency
Office of Water (4303T)
1200 Pennsylvania Avenue, NW
Washington, DC 20460

EPA 821-R-11-005

Table of Contents

Section 1: Introduction

1.1 Background

On May 13, 2005, the U.S. Environmental Protection Agency (EPA) chartered a Federal Advisory Committee on Detection and Quantitation Approaches and Uses in Clean Water Act Programs (hereafter referred to as the FACDQ or the Committee). The purpose of the Committee was to evaluate and recommend detection and quantitation procedures and uses of these procedures in Clean Water Act (CWA) programs. The Committee included 21 members representing EPA and five groups of stakeholders: laboratories, industry, publicly owned treatment works or POTWs, States, and environmental organizations. The final FACDQ meeting was held on December 21, 2007; the December 27, 2007, final FACDQ report is available at: http://water.epa.gov/scitech/methods/cwa/det/index.cfm.

Early in its work, the FACDQ reached agreement on 15 statements that described "What We Need a Procedure to Do." As part of their assessment process, the FACDQ conducted a study (the "FACDQ Pilot Study" of 2006-2007) of three detection and three quantitation procedures and used the 15 statements as criteria for evaluation of these procedures. Study results indicated that one detection and quantitation procedure included most of the elements that FACDQ members had agreed needed to be incorporated into a procedure.

Based on study results, the technical workgroup for the Committee revised the procedure to improve its performance, producing FACDQ Single Lab Procedure v2.4. Although full consensus of the Committee was not reached to support adoption of the procedure, consensus was reached that EPA should act to develop an alternative to the currently approved method detection limit (MDL) procedure, that the FACDQ's modified procedure contains elements that would be valuable to EPA in developing a new procedure, and that EPA should conduct a follow up study and formal peer review to confirm the performance of any new procedure developed and proposed by EPA. Based on these recommendations, EPA used the draft FACDQ procedure (v2.4) as a starting point for development of a new procedure, producing FACDQ Single Lab Procedure v2.4T.

Both the FACDQ v2.4 and FACDQ v2.4T procedures include separate sets of steps for calculating limits for two different types of methods. The FACDQ procedures define uncensored methods as methods for which at least 50% of method blanks yield a numerical result (regardless of detection or other reporting limits) and meet qualitative identification criteria. The FACDQ procedures define censored methods as methods for which less than 50% of the method blanks analyzed yield a numerical result (regardless of detection or other reporting limits) and meet qualitative identification criteria. The primary difference between version 2.4 and version 2.4T of the FACDQ procedure is the use of a prediction limit based on a t statistic to replace the tolerance limit (specified as "k" in the procedure) in the calculation of detection limits for uncensored methods. Version 2.4T also includes specific precision and accuracy measurement quality objectives (MQOs) which were used for this study, as well as several changes to improve the clarity and organization of the procedure.

This report describes the results of a study conducted in 2009 by EPA to evaluate both versions of the FACDQ procedure (2.4 and 2.4T) and the Lowest Concentration Minimum Reporting

Level (LCMRL) procedure. The FACDQ procedure produces data with which to calculate a detection limit (DL) and a quantitation limit (QL). EPA's Office of Ground Water and Drinking Water (OGWDW) developed the LCMRL and an associated LCMRL calculator, which produces data with which to calculate only a QL. The Committee had evaluated the LCMRL, as an alternative to the FACDQ QL, in the 2006-2007 FACDQ Pilot study. EPA designed the 2009 study to evaluate all three procedures in several laboratories using two types of analytical methods. EPA Method 200.7 represented an "uncensored" method, and EPA Method 625 represented a "censored" method. Method 200.7 measures inorganic analytes and Method 625 measures organic analytes. For the purposes of the study, laboratories were provided a copy of 2.4T and told to calculate detection limits for uncensored methods using both the k and t statistic separately, and to use the precision and accuracy MQOs specified in Version 2.4T for both sets of calculations.

1.2 Study Management and Participants

Six laboratories participated in this study, based on a competitive solicitation process, in which each interested laboratory was required to submit a pricing quote and historical data demonstrating their qualifications to perform the analytical methods used in the study. The six laboratories shown in Table 1-1 were selected to participate in this study.

Table 1-1. Participant Laboratories

Agriculture & Priority Pollutant Lab, Inc. (APPL) 908 North Temperance Avenue Clovis, CA 93611	MSE Technology Applications, Inc. 200 Technology Way Butte, MT 59702
Benchmark Analytics 477 Saucon Valley Road Center Valley, PA 18034	TestAmerica, Inc. (North Canton) 4101 Shuffel Street, NW North Canton, OH 44720
Columbia Analytical Services, Inc. 3725 E. Atlanta Avenue Phoenix, AZ 85040	TriMatrix Laboratories, Inc. 5560 Corporate Exchange Court, SE Grand Rapids, MI 49582

The primary purpose of this study was to evaluate the performance of the LCMRL procedure and versions 2.4 and 2.4T of the FACDQ procedure. While results obtained by individual laboratories were used relative to this purpose, no attempt was made to assess performance of individual laboratories. No endorsement of these laboratories is implied, nor should any be inferred. To preserve confidentiality, laboratories that participated in sample analysis were assigned numbers randomly from 1 to 6 for purposes of reporting data in the tables or lists included in this report.

Section 2: Study Objectives

The primary objective of this study was to assess whether the FACDQ Single Lab Procedure v2.4, the FACDQ Single Lab Procedure v2.4T and the modified LCMRL procedure can generate reliable estimates of the lowest concentration at which the procedure-specific measurement quality objectives (MQOs) can be achieved. This was assessed for two commonly used EPA Methods (200.7 and 625) in six laboratories (three laboratories per method). Section 2.1 describes the MQOs, and associated data quality objectives (DQOs) that were established to support this primary objective. Section 2.2 describes secondary objectives of this study. Additional details regarding the study objectives can be found in the Study Plan.

2.1 MQOs and DQOs Established to Support the Primary Study Objective

Due to differences in the nature of the FACDQ and LCMRL procedures, it was necessary to establish or identify separate MQOs that reflected the different goals of each procedure. As designed by the FACDQ, determination of detection limit (DL) and quantitation limit (QL) in the FACDQ procedure is based around MQOs for four data quality indicators: (DQIs). The DQIs being focused on in this study are:

- *False positive rate at the DL*, where a false positive is defined as concluding that the analyte is present in a sample[1] based on the DL when, in fact, it is absent. The FACDQ procedure seeks to achieve an average false rate of 1% or less; this MQO was established for both versions of the FACDQ procedure in this study.
- *False negative rate at the QL*, where a false negative is defined as concluding that the analyte is absent in a sample based on the DL when, in fact, it is present. The FACDQ procedure seeks to achieve an average false negative rate of 5%; this MQO was established for both versions of the procedure in this study.
- *Mean recovery*. Both versions of the FACDQ procedure require laboratories to achieve accuracy goals based on the mean recovery of spiked samples[1] used to determine FACDQ DL and QL values. The 2.4 version of the procedure recommends that these goals be selected based on the intended use of the analytical method. In designing this study, EPA selected separate mean recoveries MQOs for Method 200.7 and Method 625. These MQOs were documented in the 2.4T version of the procedure and applied to both versions of the procedure in EPA's study. The MQO for Method 200.7 was that the mean recovery be within 70-130%, and the MQO for Method 625 was that the mean recovery be within 40-160%.
- *Relative standard deviation (RSD)*. Both versions of the FACDQ procedure also require laboratories to achieve precision goals based on the relative standard deviation of spiked samples used to determine FACDQ DL and QL values. The 2.4 version of the procedure recommends that these goals be selected based on the intended use of the analytical method. For this study, EPA selected a 20% RSD for Method 200.7 and a 30% RSD for Method 625 as the MQOs for precision. These MQOs were documented in the 2.4T version of the procedure and applied to both methods during EPA's study.

[1] Note: Sample and spiked sample refer to reagent water or spiked reagent water for the purposes of this report.

In contrast to the FACDQ limits, which are based on four MQOs, EPA's Drinking Water Program LCMRL and its supporting Minimum Reporting Level (MRL) are determined based on a single MQO that takes into account both the precision and accuracy of the sample analysis at an adequately performing laboratory. This MQO is that the probability of the recovery of a sample spiked at a laboratory's LCMRL falling outside 50-150% should be 1%, and that the probability of the recovery of a sample spiked at the MRL falling outside 50-150% should be 1% for 75% of the population of adequately performing laboratories. When calculating the LCMRL, this MQO is assessed by fitting a regression model for measured vs. spiked concentration, and calculating 99% prediction limits around that line. Thus in this study, the MQOs are procedure-specific, with the FACDQ MQOs also varying between the two methods. Additionally, only the FACDQ procedure includes MQOs related to detection, because the LCMRL procedure is used only to develop quantitation limits, and not detection limits.

EPA notes that the MQOs established for this study are designed to support overall study objectives and failure to meet any single MQO on a particular sample did not automatically mean that the data were unacceptable for use in evaluating the DL and QL procedures.

Both the FACDQ and the LCMRL procedures include instructions to verify that laboratories can achieve the determined limits. As written, the FACDQ procedures allow laboratories to use data collected over a 12-month period to verify and, if necessary, adjust their initially-determined DLs and QLs. In the LCMRL procedure, laboratories analyze spiked samples at a multi-laboratory MRL rather than the single-laboratory LCMRLs. The results of these spike analyses are used to assess whether the laboratory is in control. Single-laboratory LCMRLs only are calculated as a step in determining the MRL, and are not used for monitoring or laboratory reporting purposes. Ongoing verification objectives were modified for both procedures to meet EPA's overall goals of this study as follows.

- Because it was impractical to verify the FACDQ limits over a 12-month period, the objective of the verification process in the study was to assess whether the procedure as a whole yields accurate estimates of the minimum concentration that met the study MQOs. To achieve this objective, the verification process was modified as described in Section 3 of this report and in the Study Plan.
- For the LCMRL, the objective was not to verify that labs could achieve a pre-determined MRL using the LCMRL procedure, but to assess whether the single-laboratory LCMRLs accurately reflect the minimum concentration to achieve the target 99% probability of a recovery falling between 50-150%. Therefore, EPA designed the verification phase to include spiking at both the laboratory's determined LCMRL and at a multi-laboratory MRL determined by EPA from those LCMRLs.

To best assess whether the FACDQ and LCMRL procedure limits accurately estimate the lowest concentration at which the procedure-specific MQOs are met, data quality objectives (DQOs) were developed to define criteria for use in assessing whether each procedure really achieves the data quality indicators it is designed to achieve. The study design (specifically, the number of laboratories and replicate analyses performed per analyte and spike level) was developed to meet study DQOs. The DQOs and associated parameters for determining whether the limits achieve their target goals are MQO and procedure-specific, and are listed in Table 2-1. As an example, for a DQI such as the false positive rate at the FACDQ DL: the MQO (or acceptance criteria) is 1% and the DQO is described by "a deviation from the target 1% false positive detection rate

that can be identified with at least 90% probability at the 95% confidence level for each method, when the actual false positive rate being achieved by the procedure is 3% or greater."

Table 2-1. Data Quality Objectives for Assessing whether FACDQ and LCMRL Limits Achieve MQOs

Procedure/Limit	Data Quality Indicator	Measurement Quality Objective (MQO)	Data Quality Objective (DQO)
LCMRL/MRL	Probability that recovery of a spiked sample falls outside 50-150%	1%	The LCMRL is designed to reflect the minimum concentration at which 50-150% of the true concentration of a target analyte will be recovered with 99% probability (i.e., it allows for a 1% failure rate for the recovery windows). The DQO is that if the actual failure rate being achieved by the procedure is 3% or greater, study data will allow EPA to identify a deviation from the 1% failure rate with at least 90% probability at the 95% confidence level. For example, if the actual failure rate being achieved by the procedure is 3.5%, the DQO is that there would be at least a 90% probability the data will demonstrate that this rate exceeds the 1% rate targeted by the procedure, and that EPA would be able to state the observed exceedance with 95% confidence.
FACDQ DL	False positive rate	1%	The FACDQ DL is designed to reflect the minimum measured concentration at which there would be a 1% false positive rate (i.e., it reflects the measured concentration that would be exceeded by no more than 1% of blank samples analyzed). The DQO is that if the actual false positive rate being achieved by the procedure is ≥3%, study data will allow EPA to identify a deviation from the 1% rate with at least 90% probability at the 95% confidence level. For example, if the actual false positive rate being achieved by the procedure is 3.5%, the DQO is that there would be at least a 90% probability that the data will demonstrate that this rate exceeds the 1% rate targeted by the procedure, and that EPA would be able to state the observed exceedance with 95% confidence.
FACDQ QL	False negative rate	5%	The FACDQ QL is designed to reflect the minimum concentration at which there would be a 5% false negative rate (i.e., no more than 5% of samples that truly contain an analyte of interest above the QL would yield measured results that are below the DL). The DQO is that if the actual false negative rate being achieved by the procedure is ≥9% or ≤2%, study data will allow EPA to identify a deviation from the 5% rate with at least 90% probability at the 95% confidence level. For example, if the actual false negative rate being achieved by the procedure is 10%, the DQO is that there would be at least a 90% probability that the data will demonstrate this rate exceeds the 5% rate targeted by the procedure, and that EPA would be able to state the observed deviation with 95% confidence. Similarly, if the actual false negative rate being achieved by the procedure is 1%, the DQO is that there would be at least a 90% probability that the data will demonstrate that this rate falls below the 5% rate targeted by the procedure, and that EPA would be able to state the observed deviation with 95% confidence.
FACDQ QL	Mean recovery (or mean bias)	70-130% or 30% mean bias (Method 200.7); 40-160% or 60% mean bias (Method 625)	The FACDQ QL is designed to reflect the minimum concentration at which the targeted mean bias (30% for Method 200.7 and 60% for Method 625) would be achieved (i.e., the FACDQ QL is designed to reflect the minimum concentration at which the mean recovery window of 70-130% can be achieved for Method 200.7 and the mean recovery window of 40-160% can be achieved for Method 625). The DQO for Method 200.7 is that if the actual mean bias being achieved by the procedure is ≥39% or ≤21%, study data will allow EPA to identify a deviation from the 30% target with at least 90% probability at the 95% confidence level. The DQO for Method 625 is that if the actual mean bias being achieved by the procedure is ≥68% or ≤52%, study data will allow EPA to identify a deviation from the 60% target with at least 90% probability at the 95% confidence level. For example, if the actual mean bias being achieved by the procedure for Method 625 is 50%, the DQO is that there would be at least a 90% probability that the data will demonstrate that this exceeds the 60% RSD targeted by the procedure, and that EPA would be able to state the observed deviation with 95% confidence.

Table 2-1. Data Quality Objectives for Assessing whether FACDQ and LCMRL Limits Achieve MQOs

Procedure/Limit	Data Quality Indicator	Measurement Quality Objective (MQO)	Data Quality Objective (DQO)
FACDQ QL	Relative Standard Deviation (RSD)	20% (Method 200.7); 30% (Method 625)	The FACDQ QL is designed to reflect the minimum concentration at which the targeted RSD (20% for Method 200.7 and 30% for Method 625) would be achieved. The DQO for Method 200.7 is that if the actual RSD being achieved by the procedure is ≥25% or ≤15%, study data will allow EPA to identify a deviation from the 20% target with at least 90% probability at the 95% confidence level. The DQO for Method 625 is that if the actual RSD being achieved by the procedure is ≥37% or ≤23%, study data will allow EPA to identify a deviation from the 30% target with at least 90% probability at the 95% confidence level. For example, if the actual RSD being achieved by the procedure for Method 200.7 is 26%, the DQO is that there would at least a 90% probability that the data will demonstrate that this exceeds the 20% RSD targeted by the procedure, and that EPA would be able to state the observed exceedance with 95% confidence.

These DQOs were developed, and the study was designed, based on the assumption that the most statistically powerful assessments of whether the limits accurately meet their target MQOs can be made on a method-specific basis, rather than on an analyte or laboratory-specific basis. In other words, all mean recoveries and RSDs calculated using samples spiked at the FACDQ QLs would be pooled to perform a single comparison to the target values, and the frequency of false positive results, false negative results and recoveries outside the LCMRL target range would be determined from data over all analytes and laboratories. This was done because combining data over all analytes within a method yields the most reliable and statistically powerful assessment.

2.2 Secondary Study Objectives

Specific secondary objectives of this study are to determine if:

1) The procedures are clearly written.
2) Data from the procedures can be easily processed in the laboratory.
3) The procedures are performed correctly by those who are expected to use them.
4) The procedures work for different types of analytical methods.

Additionally, as a secondary study objective, and in conjunction with further review of procedures, peer review and discussion with labs and stakeholders, EPA will use this study data to better determine if the procedures tested (FACDQ v2.4 and v2.4T and the LCMRL procedure) meet the 15 objectives that the FACDQ used to describe its needs for a detection/quantitation limit procedure and its resulting limits. The FACDQ identified objectives titled, "What do we need a procedure to do," are listed below:[2]

i. Does the procedure provide an explicit estimate of bias for the QL that must be verifiable by labs at those limits?

ii. Does the procedure provide an explicit estimate of precision for the QL that must be verifiable by labs at those limits? If no, does this procedure adequately estimate the bias at

[2] Report of the Federal Advisory Committee on Detection and Quantitation Approaches and Uses in Clean Water Act Programs. Appendix C. What We Need A Procedure To Do. Pages C-1 to C-6. http://water.epa.gov/scitech/methods/cwa/det/upload/final-report-200712.pdf

the QL and assess that it meets a predetermined criterion?

iii. Does the procedure provide an explicit false positive rate for the DL?

iv. Does the procedure provide an explicit false negative rate for the true value at the QL based on making the detection decision at the DL?

v. Does the procedure provide that qualitative identification criteria defined in the analytical method are met at the determined DL and QL?

vi. Does the procedure adequately represent routine variability in lab performance?

vii. Does the procedure perform on-going verification of estimates?

viii. Is the procedure capable of calculating limits using matrices other than lab reagent grade water?

ix. Does the procedure use only data that result from test methods conducted in their entirety?

x. Does the procedure explicitly adjust or account for situations where method blanks always return a non-zero result/response?

xi. Does the procedure explicitly adjust or account for situations where method blanks are intermittently contaminated?

xii. Is the procedure clearly written with enough detail so that most users can understand and implement them?

xiii. Is the procedure cost effective?

xiv. Does the procedure assess multi-laboratory and inter-laboratory variability when data from more than one lab is used?

xv. Is the procedure applicable to all users and test methods?

Section 3: Study Design and Implementation

3.1 General Description of Study Design

EPA used two analytical methods (EPA Methods 200.7 and 625) to test the FACDQ and LCMRL procedures of interest. The rationale for selecting these methods is presented in Table 3-1, along with a list of analytes targeted by each method. Historical data indicated that Method 200.7 meets the FACDQ 2.4T definition of an uncensored method, and Method 625 meets the FACDQ 2.4T definition of a censored method. Each of these methods was pre-classified accordingly for the purposes of this study (i.e., laboratories were instructed to use the FACDQ procedure for an uncensored method when using Method 200.7 and to use the FACDQ procedure for a censored method when using Method 625).

Table 3-1. Analytical Methods to be Used and Analytes to be Targeted in the FACDQ Study

Method	Rationale for Selection	Analytes to be Targeted in Each Method	
EPA Method 200.7, Trace elements via ICP-atomic emission spectroscopy (Revision 4.4)	This is a widely used multi-analyte method using optical techniques to determine metals. Detection limits for this method can be driven by blanks or instrumental sensitivity, and the method is subject to false positives.	Aluminum Antimony Arsenic Barium Beryllium Cadmium Calcium Chromium Cobalt Copper Iron Lead	Magnesium Manganese Molybdenum Nickel Potassium Selenium Silver Sodium Thallium Tin Vanadium Zinc
EPA Method 625, Capillary Column Gas Chromatography/Mass Spectrometry ***Note:*** Pesticide analytes and Aroclors are not targeted by Method 625 in this study	This is a widely used multi-analyte method using GC/MS techniques to determine semivolatile organic compounds. Detection limits for this method are often driven by qualitative identification criteria; the sample preparation stage of the method can be a source of imprecision.	Acenaphthene Acenaphthylene Anthracene Benzo(a)anthracene Benzo(b)fluoranthene Benzo(k)fluoranthene Benzo(a)pyrene Benzo(ghi)perylene Benzyl butyl phthalate bis(2-Chloroethyl)ether bis(2-Chloroethoxy)methane bis(2-Ethylhexyl)phthalate bis(2-Chloroisopropyl)ether 4-Bromophenyl phenyl ether 2-Chloronaphthalene 4-Chlorophenyl phenyl ether Chrysene Dibenzo(a,h)anthracene Di-*n*-butylphthalate Di-*n*-octylphthalate 3,3'-Dichlorobenzidine Diethyl phthalate Dimethyl phthalate 2-Nitrophenol 4-Nitrophenol	2,4-Dinitrotoluene 2,6-Dinitrotoluene Fluoranthene Fluorene Hexachlorobenzene Hexachlorobutadiene Hexachloroethane Indeno(1,2,3-cd)pyrene Isophorone Naphthalene Nitrobenzene N-Nitroso-di-*n*-propylamine Phenanthrene Pyrene 1,2,4-Trichlorobenzene 4-Chloro-3-methylphenol 2-Chlorophenol 2,4-Dichlorophenol 2,4-Dimethylphenol 2,4-Dinitrophenol 2-Methyl-4,6-dinitrophenol Pentachlorophenol Phenol 2,4,6-Trichlorophenol

A few analytes (e.g., phosphorus and titanium for Method 200.7, and Aroclors and pesticides for Method 625) listed in the methods were excluded from the list of target analytes for the study. These analytes were excluded because they are not commonly targeted by laboratories when

using these methods and therefore, laboratories are unlikely to have the existing blank data necessary for determining FACDQ DLs.

To assess whether FACDQ 2.4, FACDQ 2.4T and LCMRL generate reliable estimates of the lowest concentration at which the procedure-specific MQOs could be achieved, the study was divided into the following 3 phases, or tasks.

- **Task 1**— Compiling and submitting historical data (blank, MDL/ML and other data) and determining the "uncensored" startup FACDQ DLs for Method 200.7 based on these existing blank data. This task was divided into three subtasks that applied to one or both methods, as follows:
  - **Task 1A**: All laboratories were required to gather and submit a set of existing blank results that covered a period of approximately 6 months or 30 analytical batches, whichever yielded the greatest number of blanks (to a maximum of 100 blanks).
  - **Task 1B**: Method 200.7 laboratories then used a subset of 7 of these blanks to determine a startup FACDQ DL for each analyte using both the t- and the k- statistics (Referred to DL_T and DL_K).
  - **Task 1C**: The laboratories used a second subset of 20 of their blank results to evaluate and adjust their start-up DLs using the process described in the FACDQ procedure.

  Note: Because the blank data were only used to calculate and evaluate DLs, Task 1 applies only to the FACDQ procedure and not the LCMRL. Additionally, Task 1B only applies to Method 200.7.
- **Task 2**— Determination of FACDQ QLs and LCMRLs for Methods 200.7 and 625 and determination of FACDQ DLs for Method 625. This task was divided into three subtasks that applied to one or both methods, as follows:
  - **Task 2A**: All laboratories were required to determine the OGWDW LCMRLs by selecting seven spike levels for each analyte, preparing and analyzing four replicates at each spike level over a period of approximately two weeks with new initial calibration for Method 625 (as required by Section 9.3 of Method 200.7, calibrations were performed daily), and using the LCMRL procedure and downloadable software to calculate an LCMRL for each analyte. If an LCMRL could not be produced, a new spiking level was selected, and another four samples were prepared and analyzed at this level. A second attempt to calculate an LCMRL was made by combining data from the additional analyses with data from the initial round of analyses. Laboratories were not required to make a third attempt if an LCMRL value still could not be produced.
  - **Task 2B**: Method 200.7 laboratories used the FACDQ 2.4T procedure to determine two versions of the FACDQ QLs for each analyte. One version was based on the DL_T determined in Task 1; the other was based on the DL_K determined in Task 2. To determine these QL_{DLT} and QL_{DLK} values, each Method 200.7 laboratory followed the iterative spiking technique described in the FACDQ procedure to determine appropriate starting spike levels for each QL. Each laboratory then prepared and analyzed seven replicate samples spiked at the starting levels over a period of approximately two weeks. As required by Section 9.3 of Method 200.7, calibrations were performed daily. After completing their analyses, laboratories determined the mean recovery and RSD values for each analyte. If these values did not meet both sets of MQOs (i.e., mean recovery within 70-130% and RSD $\leq$ 20%), the labs were required to make a second attempt by preparing

and analyzing another set of seven replicate samples spiked at a new level. Laboratories were not required to make a third attempt for analytes that failed to meet the MQO. For all analytes that met study MQOs, laboratories calculated the lowest expected result (LER) and compared it to the corresponding DL determined in Task 1. If the LER was below the corresponding DL, the laboratories were required to raise the QL as described in the FACDQ procedure.

- **Task 2C**: Method 625 laboratories determined a single FACDQ QL for each target analyte by following the iterative spiking technique described in the FACDQ procedure to determine appropriate starting spike level for each analyte, preparing and analyzing seven replicates at the selected spike level over a period of approximately two weeks including a new initial calibration, and comparing the mean recovery and RSD of these analyses to the study MQOs. Each laboratory used the data generated from these analyses to estimate the FACDQ DL and the blank data compiled in Task 1 to check these values, as described in the FACDQ 2.4T procedure. If the MQOs (i.e., mean recovery between 40-160% and RSD $\leq$30%) were not met, a new spike level was selected and another set of seven replicates was prepared and analyzed over a two week period. A third attempt was not required if the MQOs were still not met. Laboratories were required to determine the LER for each analyte that met study MQOs, compare it to the corresponding DL, and follow the FACDQ procedure to adjust the LER if it was below the DL.

- **Task 3**— Ongoing verification and evaluation of FACDQ QLs and LCMRLs determined during Task 2 of the study. EPA used the results generated in Task 2 to select five spike levels for each analyte, and instructed the laboratories to prepare and analyze seven replicate samples at each of those five spike levels. The five spike levels were designed to allow EPA to determine if the FACDQ QL and LCMRL accurately estimate the minimum concentrations that meet the applicable MQOs targeted by each procedure. Accordingly, the selected spike levels varied by lab. For Method 200.7, three of the five spike levels corresponded approximately to each laboratory's LCMRL, QL_{DLT}, and QL_{DLK}. For Method 625, two of the spike levels corresponded approximately to each laboratory's LCMRL and FACDQ QL. The remaining spike levels were chosen to assess the effect of deviations from the limit calculation instructions, assure that at least one spike level is below the MRL calculated for the analyte, and to fill in any large gaps between the calculated limits. Similar to previous tasks, new initial calibrations were required for Method 625.

Analyses were conducted in a total of six labs (three laboratories per method). As noted above, laboratories were instructed to follow the appropriate sections in the FACDQ 2.4T procedure and the *Guidance Document for Determining Lowest Concentration Minimum Reporting Levels (LCMRLs)*. Labs were also provided a set of study reminders and spiking guidance. The additional spiking guidance suggested use of historical laboratory data, analyst's experience, and DL/QL data in the methods being tested.

In order to ensure temporal variability was included, laboratories were required to spread all Task 2 and Task 3 analyses over at least three different preparation/analysis batches for each concentration (i.e., there must be at least 3 preparation and 3 analysis batches associated with each spike level) over a two week period. In addition, all Method 625 laboratories were required to perform a new initial calibration in the middle of each analysis phase. (Method 200.7 requires daily instrument calibration.)

When analyzing the samples, laboratories were required to follow the analytical methods exactly as written. Deviations (other than those explicitly allowed by the flexibility included in the method) were not allowed without prior, written approval. To further minimize sources of variability, some restrictions on method flexibility were employed. For example, only continuous liquid/liquid extraction was allowed for Method 625, and acid digestion using either a hot-plate or block digester was required for all Method 200.7 samples, regardless of the turbidity. Personnel that conducted study analyses were required to be the same as those that routinely conducted analyses by the same method at the laboratory facility and who were identified in the qualification response. Laboratories also were required to carry study samples through all the same sample preparatory steps (including extraction or digestion and/or any applicable clean-up steps) and analysis steps (including instrument parameter setup) as used for typical wastewater samples. Additional information about the study design, including specific QA/QC procedures employed during the study and limitations of the study (i.e., what the study did and did not do) can be found in the Study Plan.

3.2 Deviations from the Study Design

In practice, laboratories sometimes deviated from instructions involving the frequency and amount of sample analyses. Specific instances of these deviations from the study design are described below. In other cases, laboratories deviated from requirements specified in EPA Methods 200.7 and 625; these analytical method deviations are described in Section 3.3, Data Review and Validation. Deviations from the calculations required by the procedures and/or the study plans are described in Section 3.4, Detailed Description of Limit Calculations and Data Analysis.

- Laboratory 3 did not perform any analyses for thallium because the lab does not routinely target this analyte when performing Method 200.7. This modification was approved by EPA prior to initiation of the study.
- During Task 2, all of the laboratories had different interpretations of the study requirements concerning temporal variability. After identifying this problem, EPA modified the instructions to provide additional clarity for temporal variability during the Task 3 analyses, and no further problems were observed. A discussion of the Task 2 deviations and their potential impact on study results is described in Section 5.4, Assessment of Task 2 Temporal and Batch Variability.
- When determining the LCMRL during Task 2, Laboratories 2, 3, 4, and 5 spiked and analyzed replicates at more than the required 7 concentrations for all target analytes. When re-spiking was necessary for analytes that failed to produce an LCMRL based on the initial set of analyses, Laboratories 1 and 3 analyzed samples at two new spike levels instead of the one new spike level required. Details on how EPA utilized these data are described in Section 5.1, LCMRL Determinations.
- When it was necessary to re-spike and re-analyze 7 replicates at a new concentration for analytes that did not meet the FACDQ procedure MQOs, Laboratory 2 analyzed 21 replicates at the new concentration instead of the required 7 (the lab ran 3 batches each containing 7 replicates). When contacted about this, the lab indicated that they had been confused while moving back and forth between the various requirements for the LCMRL and the FACDQ

procedures. Details on how EPA utilized these data are described in Section 5.1, LCMRL Determinations.

- Due to data reporting errors, Laboratories 5 and 6 each failed to re-spike and re-analyze samples for one analyte (butyl benzyl phthalate for Laboratory 5, and 3,3'-dichlorobenzidine for Laboratory 6) that failed to meet FACDQ procedure MQOs during Task 2. The error was not identified until this phase of the study was completed.
- Many laboratories used spiking solutions that contained all analytes, rather than customizing specific solutions that contained only the subset of analytes that were intended to be targeted at that spike level. This was not a true deviation from the study design, and it reflected strategies that laboratories would be likely to employ should the procedures be adopted. It also yielded a benefit of providing additional data that could be used to offset actual deviations from the study design. For example, Lab 5 initially selected a spike level for di-n-butyl phthalate that was too low based on their maximum historical blank data, but because the lab had a second data set at a higher spike level, the lab was able to use the di-n-butyl phthalate data from the higher level to determine a QL that met the FACDQ 2.4T procedure requirements.
- Laboratory 3 did not perform downspiking analyses, due to a misunderstanding of the instructions. Instead the laboratory spiked replicate samples at multiple levels. The other five laboratories did perform downspiking analyses, but did not always follow the exact downspiking instructions in the procedures and the relationship between the downspiking results and the initial QL spike levels chosen for each analyte was unclear. This issue is addressed further in Section 5.2.3, Assessment of FACDQ QL Downspiking Analyses,.
- When adjusting the QL for any analytes in which the LER was below the corresponding DL (as required by the FACDQ procedure), several of the labs that used spiking solutions that contained all analytes, provided data from all spike levels for the QL spikes used to determine the FACDQ limits. As a result, some of the calculations, including the LER check and DL calculation, were performed on multiple spike levels, and it was not clear which spike level was the initial spike level chosen by the laboratory based on the procedure and downspiking results. In part this was due to an unanticipated limitation of the data reporting format and the tendency to copy and apply formulas to all sets of data. EPA was able to determine the most appropriate starting spike levels, DLs, and QLs for each analyte based on the data, and the appropriate data were used for all data analyses described in this report. Therefore, these deviations had no impact on the study results.
- When determining and evaluating LERs in Task 2, Lab 5 sometimes compared them to the wrong DL value (i.e., the DL associated with a different spike level than the one that should have been used.) EPA was able to perform the correct calculation and comparison for each analyte, and the correct values were used for all data analyses in this report. Therefore, this Lab 5 deviation had no impact on study results.
- Although Lab 6 correctly followed the downspiking procedure described FACDQ v2.4T, the lab initially selected an inappropriately low initial spike level for one analyte. (The selected value was less than two times their maximum Task 1 blank value as required by the FACDQ procedure and the study design.) The problem was not due to misunderstanding; it stemmed from an incorrect linkage between their Task 1 blank data and their Task 2 calculation files. The analyte in question was analyzed at a higher spike level during the re-spiking phase of

Task 2, and a QL that reflected the FACDQ procedure spiking requirements was produced. Therefore, this deviation had no impact on study results.

- One of the Task 3 samples analyzed by Laboratory 5 was confirmed to be spiked at an incorrect level, based on further investigation with the laboratory. Because this sample did not provide acceptable results for any analyte, only six replicates were available from Laboratory 5 for their lowest spiking level in Task 3.
- One of the Task 3 samples analyzed by Laboratory 6 not properly acidified during the preparation step, resulting in low recoveries of the acid-extractable compounds. The laboratory re-prepared and reanalyzed the sample, but in doing so, spiked it at the wrong level. Since neither analysis for this sample provided acceptable results, only six replicates were available from Lab 6 for their fifth spiking level in Task 3.

3.3 Data Review and Validation

The data from all six laboratories was reviewed and validated as soon as possible after receipt. Data packages included electronic summary level data for study and QC samples, calculation spreadsheets, and raw data. Raw data included calibration data, chromatograms, quantitation reports, spectra, bench sheets, and laboratory notebooks showing weights, volumes, manual calculations, and other data that would allow verification of the calculations performed and would allow all final results to be traced to the raw data. Data were reviewed against requirements in Methods 200.7 and 625 and the Study Plan to ensure that: results from each laboratory were complete (i.e., that all required data were present); all samples were analyzed properly; appropriate spike levels were used; the analytical systems were properly calibrated; and results calculation procedures were followed correctly. In cases where a laboratory deviated from some of the electronic data reporting format instructions in their initial submissions, EPA worked with that laboratory to resolve the problems to ensure correct interpretation and analysis of data and facilitate the data review. None of the identified problems with the data reporting format were significant or affected data quality. EPA also spot checked the raw data to verify use of proper calculations, verify that electronic results reflected raw data results, and that qualitative identification criteria were met when reporting positive results. Standardized data review checklists were used to facilitate and document these activities.

A fundamental objective of this review was to maximize data use, and laboratories were contacted to resolve questions and/or discrepancies. Deviations from the study design requirements were documented in Section 3.2 of this report. As noted in Section 3.2, most of those deviations could be corrected through discussions with the laboratories. Deviations from analytical method requirements were as follows. None of these was considered significant enough to warrant exclusion of data from the study data set.

- Several labs reported sample results were slightly outside calibration ranges. In each case, results in the study database were flagged with either "REXC" for results that exceeded the calibration range or "RBC" for results that were below the lower limit of the calibration range. Generally, this situation was observed for samples that were spiked at or near the upper or lower limit of the range, and recoveries of these samples were generally within acceptable ranges.

- Method 625 requires that the response for any analyte in the calibration verification standard be within 20% of the predicted response. Method 200.7 requires that recoveries from analysis of calibration verification standard be between 90-110%. Several laboratories reported a few calibration verification standard results that were slightly outside these criteria. Results that exceeded the applicable criterion were flagged with "HVER" in the study database, and results that fell below the Method 200.7 criterion were flagged with "LVER" in the database.
- Method 200.7 requires that laboratories analyze a calibration blank immediately after analyzing the calibration standard. Laboratory 3 analyzed their calibration blanks prior to each verification standard instead of after. The laboratory analyzed the correct number of calibration blanks, and all samples were bracketed by calibration blanks analyzed before and after the sample set. Therefore, sample results are unlikely to be affected by this minor deviation.
- Due to the nature of this study, normally acceptable thresholds for blank contamination were not applicable. Therefore, laboratories were required to ensure that their preparation and/or equipment blanks that were as free from contamination as possible. For the purposes of evaluating blanks, EPA compared each laboratory's study blanks to the same laboratory's historical blank data. Preparation blank results that exceeded the historical blank values were flagged with "PB" in the database; calibration blank results that exceeded the historical blank values were flagged with "CB" in the database.

3.4 Detailed Description of Limit Calculations and Data Analysis

All limit calculations and MQO assessments performed during Tasks 1 and 2 were initially performed by the analytical laboratories. The calculations were reviewed and where necessary were revised by EPA to ensure consistency with the procedures and comparability between laboratory results. EPA performed additional calculations during Task 2 to determine multi-laboratory MRLs and revise LCMRLs based on changes to the procedure algorithm. All Task 3 calculations and data analyses were performed under EPA oversight. Limit calculations and data analyses are presented by task in the Sections 3.4.1-3.4.3.

3.4.1 Task 1 Limit Calculations

For Method 200.7, Task 1 included steps to calculate startup DLs (Task 1B) and perform ongoing verification on the startup DLs using a subset of the remaining blanks (Task 1C). In Task 1B, laboratories calculated a startup DL_T by using the 7 oldest blank results per analyte and the DL formula given in the FACDQ 2.4T procedure. The laboratories also calculated a startup DL_K using the tolerance limit, k, and the formula given below:

$$DL_K = \overline{X} + sK_{(n-1,0.99,0.01)}$$

Where:

- $\overline{X}$ *is the mean result from the method blanks*
- $K_{(n-1,0.99,0.01)}$ *is a multiplier for a tolerance limit based on 99% coverage probability of 99% of the population of routine blanks and n-1 degrees of freedom.*

In Task 1C, the laboratories performed ongoing verification of the startup DL_T and DL_K described above using the next oldest 20 blank results. This was done by:

1) Following Section 1.5 of the FACDQ 2.4T procedure to identify any outlying results from these blank data, and removing a maximum of one outlier per analyte/lab set.
2) Performing the ongoing verification blank check given in Section 5.3 of the FACDQ 2.4T procedure to adjust the DL where necessary.

Task 1B did not apply to Method 625 because the FACDQ DL was to be calculated using spiked sample data analyzed during Task 2. However, the outlier test and identification of the highest non-outlying blank was performed for this method using the 20 oldest blank results identified during Task 1C.

All Task 1 calculations were performed by the laboratories and submitted for review. In cases where the calculations did not follow the procedure and/or study plan, the calculations were corrected accordingly, and provided to the laboratory with the corrected limits. Specific issues regarding Task 1 calculations that required correction are listed below:

- When calculating DL_T and DL_K values for Method 200.7, Laboratories 1 and 2 used negative mean values instead of replacing these values with zero as instructed by the FACDQ procedure.
- Laboratory 1 failed to identify outliers for four of the analytes. Also, instead of comparing the second subset of blanks to the DL values determined with the first subset of blanks, Laboratory 1 recalculated their DL values using the second set of blanks.
- Laboratory 2 did not use the second set of blank results to adjust their DL_T values, but did do so for their DL_K values.
- Laboratory 3 performed an outlier test using the mean and standard deviation calculated from the Task 1B blanks (i.e., those used to calculate the startup DL), rather than the mean and standard deviation of the Task 1C blanks (i.e., those used in the ongoing verification blank check).
- When performing outlier tests on their historical (Task 1) blanks, Laboratory 4 did not replace their non-detect results with zero as required by the FACDQ procedure.

3.4.2 Task 2 Limit Calculations

Task 2 included determination of the LCMRL (Task 2A), the FACDQ QL for Method 200.7 (Task 2B), and the FACDQ DL and QL for Method 625 (Task 2C). LCMRL calculations are described in Section 3.4.2.1, and FACDQ procedure limit calculations are described in Section 3.4.2.2.

3.4.2.1 LCMRL Calculations (Task 2A)

During Task 2A, laboratories performed the LCMRL procedure calculations for each analyte. Calculations were performed using automated software designed specifically for this purpose. As described in Section 2, the LCMRL is an estimate of the minimum concentration at which 50-150% recovery can be achieved in a single sample with 99% probability. This estimate is made

based on fitting a linear or nonlinear model of measured concentration vs. true concentration across the range of spike levels assessed in the study. Prediction limits are then fit around the model, and the highest concentration at which these bounds intersect with 50% or 150% recovery is identified.

The LCMRL calculations were performed by the laboratory and submitted for review. The calculations were reviewed and, where necessary, such as when more data were used than were specified by the Study Plan, recalculated the limits.

EPA used the individual laboratories' LCMRLs to determine a multi-laboratory MRL for each analyte. Additionally, EPA recalculated the individual LCMRLs, due to a recent modification to the calculation algorithm. This change is discussed in Section 5.1.3.

3.4.2.2 Task 2 FACDQ QL/DL Calculations (Task 2B, 2C)

For both methods, laboratories prepared and analyzed seven replicates at a spike level chosen based on the downspiking criteria given in the FACDQ 2.4T procedure. (These criteria include using the results of the Task 1 calculations.) Once this was done, the laboratories calculated the mean recovery and RSD and compared them to the method and procedure-specific MQOs. If either the calculated mean recovery or the RSD MQO was not met for a given analyte, the laboratories spiked and analyzed seven replicates at a higher concentration, and performed the MQO calculations on these data. The FACDQ QL was set to the lowest concentration that met the mean recovery and RSD MQOs for the given analyte. Laboratories then performed the LER calculations given in Section 4 of the FACDQ 2.4T procedure, and adjusted the QL based on the LER where necessary.

For Method 625, laboratories also calculated the FACDQ DL using the initial spike FACDQ QL spike level results, or the FACDQ QL respike results (in cases where not all replicates at the initial level yielded a measurable signal or met qualitative identification criteria). The calculated DL then was compared to the highest non-outlying blank determined in Task 1, and adjusted when that blank value exceeded the DL.

EPA reviewed the laboratory calculations and where necessary, recalculated the limits and provided the revised limits to the laboratory. Specific issues regarding Task 2B and 2C calculations are listed below:

- Laboratory 4 adjusted the QLs based on the calculated LER regardless of whether the LER assessment passed or failed, resulting in incorrect QLs for analytes for which the LER assessment had passed.
- Laboratory 5 did not include any QL adjustment calculations in their Task 2 spreadsheet. However, this had no impact on the laboratory's results because the LER always exceeded the DL when their MQOs were met.
- When determining and evaluating LERs in Task 2, Laboratory 5 sometimes compared them to the wrong DL value (i.e., the DL associated with a different spike level than the one that should have been used.) EPA was able to perform the correct calculation and comparison for each analyte, and the correct values were used for all data analyses in this report. Therefore, this deviation had no impact on study results.

- Although Laboratory 6 correctly followed the downspiking procedure described FACDQ v2.4T, the lab initially selected an inappropriately low initial spike level for one analyte, due to an incorrect linkage between their Task 1 blank data and their Task 2 calculation files. The analyte in question was analyzed at a higher spike level during the re-spiking phase of Task 2, and a QL that reflected the FACDQ procedure spiking requirements was produced. Therefore, this deviation had no impact on study results.

3.4.3 Task 3 Calculations and Analyses

Based on the limits determined in Task 2, five different spike levels were chosen for each analyte/lab set, and laboratories spiked and analyzed seven replicates at each of those concentrations. These data were used to assess whether the FACDQ QLs and LCMRLs determined in Task 2 accurately estimated the minimum concentration to achieve the procedure and method-specific MQOs. Additionally, remaining blank data submitted as part of Task 1 (and QC blanks submitted during Tasks 2 and 3 where necessary to increase statistical power) were used to assess whether the FACDQ DLs determined in Tasks 1 or 2 accurately estimated their target MQO.

The MQO assessments performed included the following:

- Determining the frequency of blank results exceeding the FACDQ DLs using the remaining existing blank data, and comparing this frequency to the target 1% rate.
- Determining the frequency of spiked sample results below the FACDQ DLs at the Task 3 spike level(s) corresponding to the FACDQ QL, and comparing this frequency to the target 5% rate.
- Determining the frequency of spiked sample results with recovery outside of 50-150% at the Task 3 spike level corresponding to the LCMRL, and comparing this frequency to the 1% rate targeted by the LCMRL procedure.
- Determining the mean recovery and RSD at the Task 3 spike level corresponding to the FACDQ QL(s), and comparing them to the method-specific MQOs listed in the FACDQ procedures.

The frequencies, mean recoveries and RSDs described above were pooled over all laboratories and analytes for each method. This was done because analyte and laboratory-specific assessments were not of interest, and because combining this information gave the most reliable and statistically powerful assessment of the procedures. However, because the FACDQ QL includes multiple target MQOs, it could not always be assumed that the limit will always be an estimate of the minimum concentration to achieve each of these criteria for every analyte and laboratory. Therefore, the Task 2 calculations and data were assessed to identify, where possible, which MQO criterion was "limiting," i.e., which of the MQO criteria was achieved at the highest concentration, for each analyte/lab set. Once this was done, the mean recoveries, RSDs, and false negative frequencies were pooled across only those analyte/lab sets for which that MQO was determined to be limiting. The comparison of the pooled values to the target MQO was performed for all analyte/lab sets, and for the subset of analyte/lab sets for which that MQO was determined to be limiting.

The pooled false positive, false negative and LCMRL recovery criterion failure frequencies were compared to the target rates, using a one-sample proportion test following the Binomial distribution. Because a large number of blank results were available for each analyte/lab set, the effect of statistical dependence between analytes for a single blank was mitigated through the use of a Bootstrap estimation technique. These calculations are described in Section 4 and 6 of this report. The LCMRL recovery and false negative rate assessments are presented in Section 7 of this report. All of these tests were performed at the 95% confidence level.

Mean recovery and pooled RSDs were compared to the FACDQ procedure MQO criteria using a combination of parametric one-sample t-tests and nonparametric signed-rank tests, depending on the distributions of the calculated values. All of these tests were performed at the 95% confidence level. The results of these tests are presented in Section 7.

Additional analyses also were performed to assess why any departures from the limits' targets occurred. These included assessments of the effect of temporal variability, departures from the procedures' instructions, the effect of the censored/uncensored classification for Method 625, and the effect of alternate calculations or approaches suggested during a peer review of the FACDQ 2.4T procedure. These analyses included performing the MQO assessments based on different subsets of the data, F-tests to compare the variability of data covering various time periods, and fitting models of Task 3 data across multiple concentrations. These additional analyses are described and presented in Sections 4, 6, and 7.

Additional calculations were performed by the analytical laboratories during Task 3. These calculations were used to assess the clarity and ease of calculations included in the evaluated procedures, and were not used in the assessments of the limits themselves. Although these calculation errors did not directly affect the study analyses, they are listed below:

- When calculating LERs in Task 3, Laboratory 3 correctly compared the LER to DL_T but instead of also comparing the LER to DL_K, Lab 3 compared the DL_K to DL_T. Laboratory 1 calculated the LERs correctly, but did not compare them to the DLs. Laboratories 4 and 6 compared their LER values to a DL based on Task 3 data instead of the appropriate DL from Task 2. Lab 6 also did not adjust the LER when the Task 3 spike level did not match the FACDQ QL sent to the laboratory.
- When calculating the Prediction Interval of Results (PIR) in Task 3, Laboratories 1, 2, and 5 used the incorrect t-statistic (the 95th percentile t-statistic was used instead of the 99.5th percentile statistic). Laboratory 5 also adjusted the mean by the interval width twice, such that the upper and lower bound were too high. Laboratory 4 used the appropriate t-statistic, but applied it to the RSD rather than the standard deviation of the recoveries. Laboratory 6 used an incorrect square root value— the square root of (1 +1/6) instead of (1+1/7) —when calculating the PIR. Laboratory 3 used the correct values when calculating the PIRs, but did not compare them to the 50-150% window as required. Instead, Lab 3 compared the mean to the calculated PIR limits.

Section 4: DL Assessment – Method 200.7

As described in Section 3.4, detection limits were calculated for each analyte/lab set according to the steps and formula in the FACDQ 2.4T procedure. In addition to these limits (DL_T), detection limits also were determined using a tolerance limit for 99% coverage and 99% probability (DL_K). A discussion of how various steps in the FACDQ 2.4T procedure affected the resulting limits is presented in Section 4.1.

Although DL_T is based on an approximate prediction limit and DL_K is based on a tolerance limit, both detection limits were described as targeting a false positive rate of less than or equal to 1% in the versions of the FACDQ procedure in which they were presented. Therefore, EPA compared results generated for both sets of limits to the target 1% false positive rate. This comparison is presented in Section 4.2. Assessments of how various assumptions made in the procedure, including the assumption of blanks following a normal distribution, the relationship between short-term and long-term variability, and the effect of outlier testing on the resulting limits, affect the resulting rate are presented in Section 4.3.

4.1 DL Calculations

4.1.1 Differences between DL_T and DL_K

When based on seven replicates, the value for k (a 99% coverage/99% probability tolerance limit) is 6.101, compared to the 99th percentile t-statistic of 3.14. Therefore, assuming that the mean blank is close to 0 (or a negative value replaced with 0 according to the FACDQ 2.4T procedure) the resulting DL_K would be 1.94 times greater than the corresponding DL_T. When the mean blank is greater than 0, or when the blank evaluation in Task 1C results in an adjustment of one or more of the DLs, this ratio will be smaller.[3]

To evaluate the relationship between DL_T and DL_K, EPA calculated a ratio of the DL_K compared to the corresponding DL_T for each of the 71 analyte/ laboratory sets generated during Task 1 of the Procedure Evaluation Study.[4] The median and geometric mean of these ratios were 1.87 and 1.73, respectively. For 29 analyte/laboratory sets, the ratio was the maximum possible value of 1.94 (i.e., where the mean blank was set to 0, and neither DL was adjusted in Task 1C), and for 5 analyte/laboratory combinations, the ratio was the minimum possible value of 1 (i.e., where both DL_T and DL_K were adjusted to the same maximum blank value in Task 1C).
Table 4-1 shows the number of adjustments made to DL_T and DL_K based on the 20 Task 1C blanks for each Method 200.7 analyte. Overall, DL_T was adjusted for 21.1% of the analyte/lab combinations, and DL_K was adjusted for 7% of the combinations. Because Task 1C is intended to simulate the ongoing verification steps in the FACDQ 2.4T procedure, this indicates that these

[3] Task 1C required laboratories to evaluate, and if appropriate, adjust their DL as described in Section 2.3 of the FACDQ 2.4T procedure. Briefly, the procedure specifies that if 5% or more of the blank results remaining after outlier removal were greater than the DL, then the DL was to be adjusted (raised) to one of the following: the highest result if fewer than 30 blanks were available; the next highest result if 30-100 blanks were available; or to the level exceeded by 1% of the blanks if more than 100 blanks were available. Task 1C specified the use of 20 blanks, so any necessary adjustments were made to the highest result after outlier removal.
[4] As noted in Section 3.4, EPA verified, and where necessary, corrected each laboratory's limit calculations and performing statistical analyses of data.

steps have a larger effect when the DL is calculated using the t-statistic compared to when it is calculated using the tolerance limit k.

Table 4-1. Frequency of Task 1C DL Adjustments per Analyte, Method 200.7

Analyte	# DL Adjustments Based on Task 1C	
	DL_T	DL_K
Aluminum	0	0
Antimony	0	0
Arsenic	0	0
Barium	2	0
Beryllium	0	0
Cadmium	0	0
Calcium	1	0
Chromium	1	1
Cobalt	0	0
Copper	0	0
Iron	1	0
Lead	0	0
Magnesium	1	1
Manganese	1	1
Molybdenum	0	0
Nickel	1	0
Potassium	2	0
Selenium	1	0
Silver	0	0
Sodium	1	1
Thallium	1	0
Tin	0	0
Vanadium	0	0
Zinc	2	1
Total Adjustments	**15**	**5**
Percent of Limits Adjusted	**21.13**	**7.04**

4.1.2 Outlier Removal

Because application of an outlier test to blank results is part of the FACDQ 2.4T procedure, each laboratory was required to apply this test to the 20 Task 1C blanks. Because the procedure permits removing a maximum of 5% of total blank results per analyte, only one Task 1C blank could be removed from each set of 20 blanks. Table 4-2 shows the frequency of outlier removal per analyte for all 60 blanks. Table 4-2 also shows how frequently outlier removal changed the calculated DL, either because not removing the outlier would have affected whether the Task 1B DL was raised or the extent to which it was raised.

An outlier was identified and removed from the Task 1C blank data for 57 of the 71 analyte/lab sets after applying the outlier test described in the FACDQ 2.4T procedure. The removal of 21

of these 57 outliers affected the resulting DL_T, and removal of 15 of the 57 outliers affected the resulting DL_K.

Table 4-2. Frequency of Outlier Removals and Impact on Resulting DL values

Analyte	# Outliers Removed (3 maximum for 60 blanks)	# Times Outlier Removal Affected Adjustment	
		DL_T	DL_K
Aluminum	3	1	0
Antimony	0	0	0
Arsenic	2	0	0
Barium	3	1	1
Beryllium	3	1	1
Cadmium	3	1	1
Calcium	3	1	1
Chromium	3	1	1
Cobalt	2	0	0
Copper	2	1	0
Iron	3	3	2
Lead	1	0	0
Magnesium	3	1	1
Manganese	3	2	2
Molybdenum	2	0	0
Nickel	3	2	1
Potassium	3	2	1
Selenium	2	0	0
Silver	3	1	0
Sodium	2	1	1
Thallium	1	0	0
Tin	3	0	0
Vanadium	3	0	0
Zinc	2	2	2
Total	**57**	**21**	**15**
%	**80.28**	**29.58**	**21.13**

The results presented above were based on the outlier test specified in the FACDQ 2.4T procedure. An assessment of other types of outlier tests on the resulting DLs is presented in Section 4.3.2.

4.2 MQO Assessment

The number of blank results per analyte submitted by the three participant laboratories ranged between 78 and 131. Among these blanks, 7 were used to calculate an initial DL in Task 1B, and 20 others were used to verify and modify the DL in Task 1C. Because these blanks were used in the DL calculation, any comparison between these results and the calculated limit would be biased. Therefore, between 51 and 104 blanks per analyte and laboratory remained to assess the

calculated DLs. This set of blank results is referred to in this section as the "full verification set."

EPA compared the full verification set of blanks to the calculated DL_T and DL_K values for the given analyte/ laboratory set; any blank result exceeding the DL was categorized as a false positive result for that limit. The percentage of false positives over all analytes and laboratories was then calculated for each of the two DLs. The overall false positive percentages for the two limits are presented in Table 4-3, and displayed for each analyte (over all three labs) in Figure 4-1. EPA also calculated the false positive rates after applying the FACDQ 2.4T procedure's outlier removal test to the full verification set data; the rates per analyte after outlier removal are displayed in Table 4-3 and Figure 4-2.

Table 4-3. Method 200.7 False Positive Rates, Full Verification Dataset

Outliers Removed in Full Verification Data	Limit	# Total Blank Results	# False Positives	% False Positives
No	DL_T	5,032	172	3.42
	DL_K		57	1.13
Yes	DL_T	4,845	100	2.06
	DL_K		20	0.41

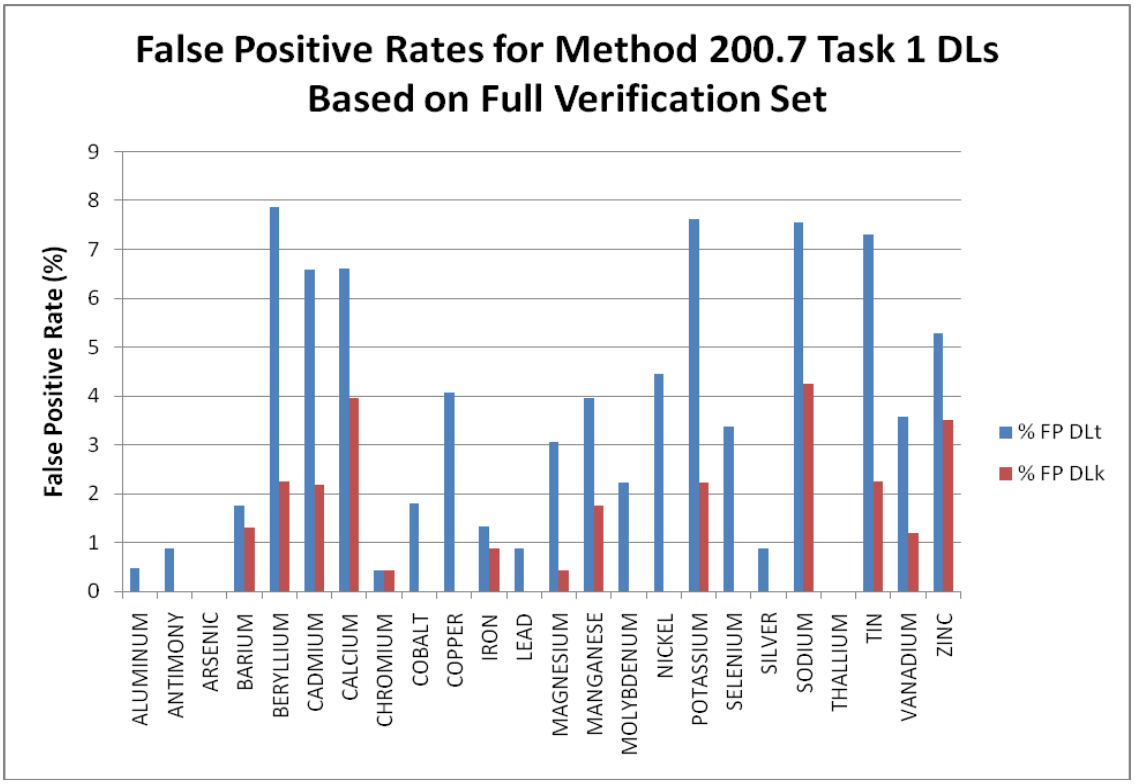

Figure 4-1. False Positive Rates Estimated for Each Analyte – Full Verification Dataset

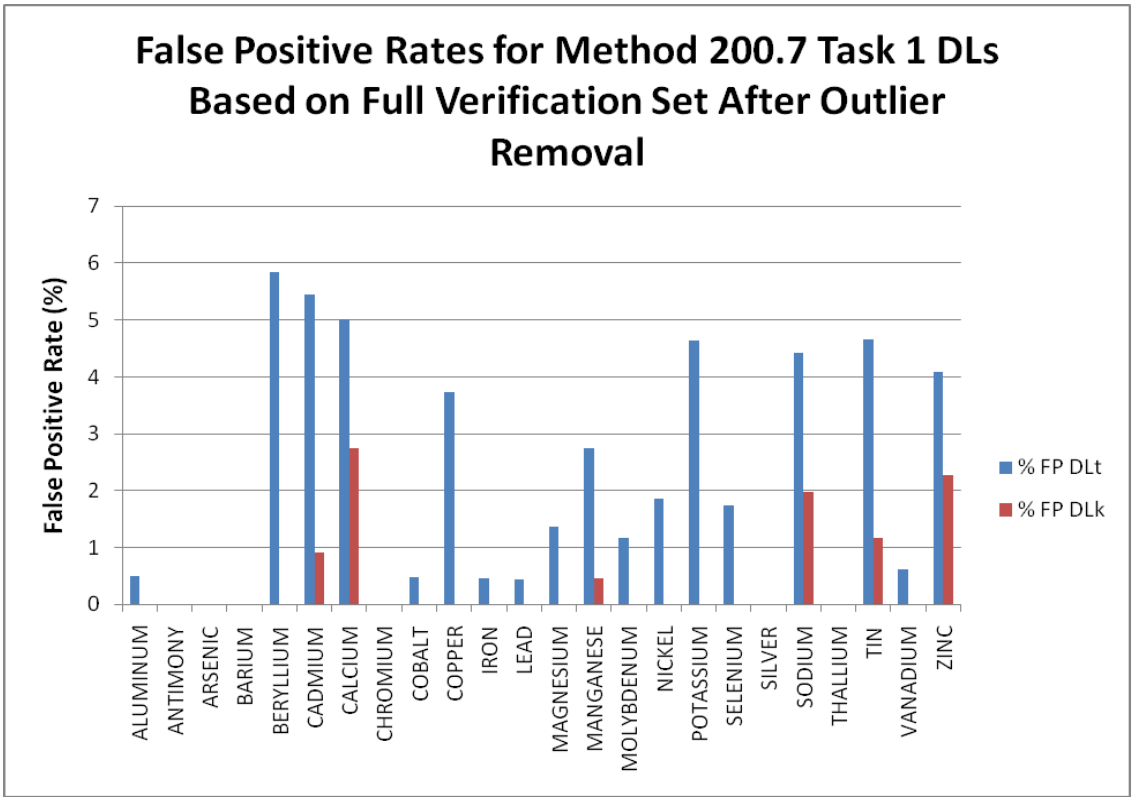

Figure 4-2. False Positive Rates Estimated for Each Analyte – Full Verification Dataset After Outlier Removal

The false positive rates presented in the above table and figures could be biased because they were calculated using all blank results and, therefore, factors influencing whether a blank result is low or high would likely affect more than one analyte. However, the large number of blank results per analyte in the full verification set allows an assessment to be made using a randomly selected subset of the data. Therefore, EPA determined a bootstrap estimate of the false positive rate by randomly selecting 100 sets of 20 blank results with replacement for each analyte/lab set. The results selected for one analyte were not from the same 20 blanks selected for each of the other analytes, thereby mitigating the effect of analyte correlation on the statistical comparison. For consistency, EPA used the same randomly selected subsets to assess DL_T and DL_K within each bootstrap run. The mean of the 100 false positive rates calculated over all analyte/lab sets for each bootstrap run was the bootstrap estimate of the false positive rate for that limit. Once the bootstrap estimates were calculated, they were compared to the target 1% false positive rate using a one-sample proportion test following the Binomial distribution, run at the 95% confidence level.

The bootstrap-estimated false positive rates for the two limits and the result of the proportion test are presented in Table 4-4. Bootstrap estimates for the false positive rates also were calculated for each analyte with and without outlier removal, and are presented in Figures 4-3 and 4-4, respectively.

Table 4-4. Method 200.7 False Positive Rates, Bootstrap Estimates

Outliers Removed in Verification Data	Limit	# Total Blanks per Bootstrap Run	Mean False Positive Rate	p-value for Proportion test
No	DL_T	1,420	3.09	<0.001
	DL_K		0.94	0.44
Yes	DL_T	1,420	1.86	0.0015
	DL_K		0.34	0.0047

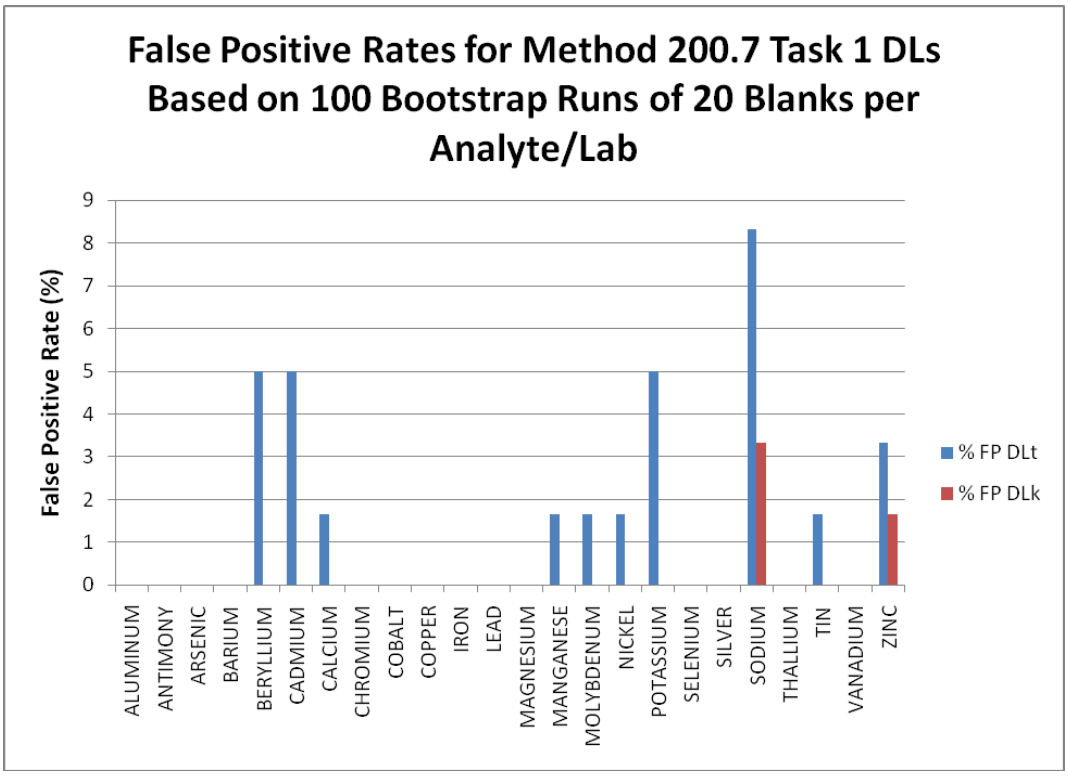

Figure 4-3. False Positive Rates Estimated for Each Analyte, Based on Bootstrap Estimation

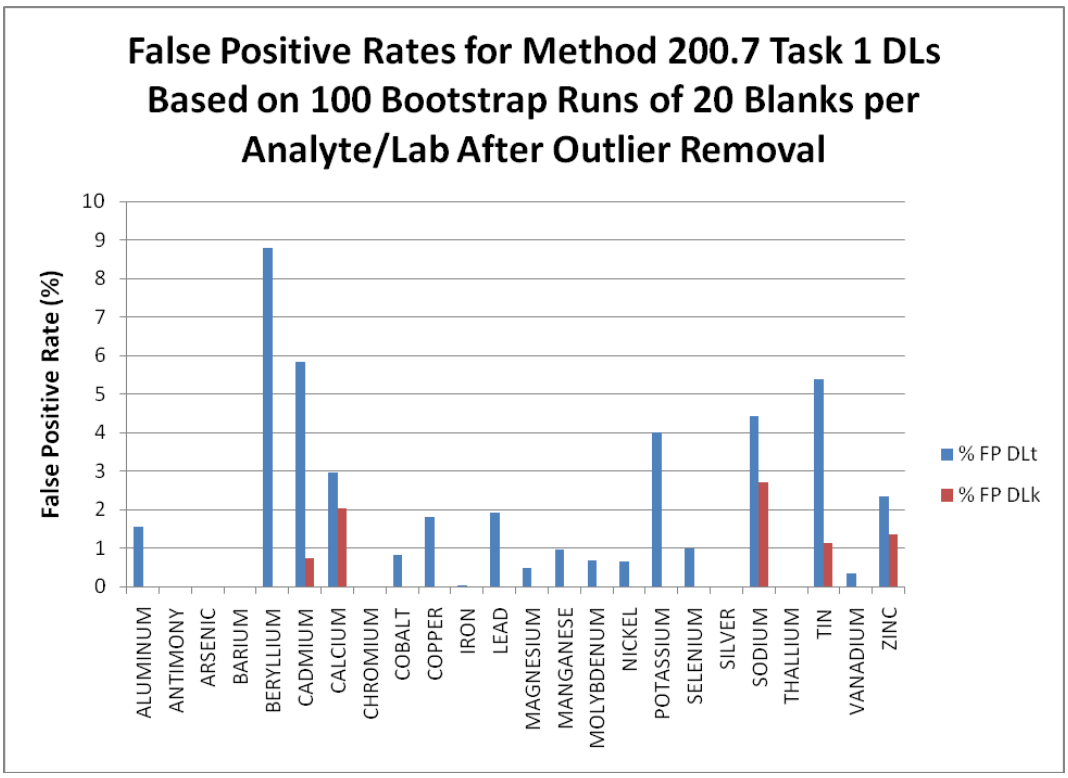

Figure 4-4. False Positive Rates Estimated for Each Analyte, Based on Bootstrap Estimation After Outlier Removal

At the 95% confidence level, the DL_T calculation yielded a false positive rate significantly greater than the 1% target, but the DL_K calculation yielded a false positive rate that was not significantly different from the 1% target even though the DL_T targets a 1% false positive rate on average, while DL_K targets a limit that would give a false positive rate at or below 1% with 99% probability. In fact, as shown in Figure 4-2, the false positive rate for DL_K exceeds 1% for 4 of the 24 analytes. As a result, although DL_K achieves the FACDQ procedure false positive rate MQO more effectively than DL_T, it does not appear to achieve the intended goal of a 99% coverage/99% probability tolerance limit. After outlier removal, the false positive rate for DL_T was significantly greater than 1% while the false positive rate for DL_K was significantly lower than 1%.

To help understand the deviations of the DLs from their target MQOs in the FACDQ procedure, EPA conducted an assessment of the MQO assumptions and the results are discussed in the following sections.

4.3 Factors Affecting MQO Assessment

4.3.1 Distribution of Blanks

The ability of the t and k multipliers to yield detection limits that meet the target false positive rate MQOs is based on the assumption that the blanks follow an approximately normal distribution. However, strong departures from a normal distribution could still result in limits that do not meet the target 1% false positive rate.

For each analyte and laboratory, EPA used the the D'Agostino Omnibus test for normality to assess the distribution of all blank results (including all blanks used in the calculation and verification of the DLs as well as the full verification set described in the previous section). This test is sensitive to departures from normality due to skewness (a non-symmetric distribution) and kurtosis (a distribution with a higher peak), and is generally considered to be one of the most powerful tests for assessing the normality assumption. In addition to the omnibus test, there are separate D'Agostino tests for levels of skewness and kurtosis that are significantly different from what is expected under a normal distribution. For analytes/laboratories for which the omnibus test was significant, EPA also performed the separate skewness and kurtosis tests to identify the nature of the departure from a normal distribution. Table 4-5 presents the results of these normality tests.

Table 4-5. Results of D'Agostino Normality Tests

Analyte	# of Sets Passing Normality Test	Lab 1		Lab 2		Lab 3	
		Test Result	Departure Type[1]	Test Result	Departure Type[1]	Test Result	Departure Type[1]
Aluminum	1	Fail	-S, +K	fail	-S	pass	
Antimony	3	Pass		pass		pass	
Arsenic	2	fail[2]		pass		pass	
Barium	0	Fail	-S, +K	fail	+S, +K	fail	+S, +K
Beryllium	1	Pass		fail	+K	fail	+S, +K
Cadmium	2	Pass		pass		fail	+S, +K
Calcium	0	Fail	+S,+K	fail	+S, +K	fail	-S, +K
Chromium	0	Fail	-S, +K	fail	-K	fail	+S, +K
Cobalt	0	Fail	+K	fail	-K	fail	+S
Copper	3	Pass		pass		pass	
Iron	0	Fail	+S, +K	fail	+S, +K	fail[2]	
Lead	3	Pass		pass		pass	
Magnesium	1	Fail	+S, +K	fail	+S, +K	pass	
Manganese	1	Fail	+K	pass		fail	+S, +K
Molybdenum	0	Fail	+K	fail	+S	fail	+S
Nickel	1	Fail	+K	fail	+S	pass	
Potassium	0	Fail	+S, +K	fail	+S, +K	fail	+S, +K
Selenium	2	Pass		pass		fail	+S
Silver	1	Fail	+S,+K	pass		fail	+K
Sodium	0	Fail	+S, +K	fail	+S, +K	fail	+S
Thallium	1	Fail	-S, +K	pass		N/A	N/A
Tin	2	Fail	+S,+K	pass		pass	
Vanadium	0	Fail	+S	fail	+K	fail	+S, +K
Zinc	0	Fail	+S, +K	fail	+S, +K	fail	+S, +K

[1] +S: positive skewness, -S: negative skewness, +K: positive kurtosis, -K: negative kurtosis
[2] D'Agostino omnibus test failed, but D'Agostino skewness and kurtosis tests each passed

Only three analytes passed the D'Agostino omnibus normality test for all three laboratories: antimony, copper, and lead. Four analytes (arsenic, cadmium, selenium, and tin) passed for two of the three laboratories. Ten analytes did not pass the D'Agostino omnibus test for any of the three laboratories. Generally, the analytes that more frequently passed the normality test tended to be analytes that are more heavily regulated, or are regulated to lower levels, than those that

tended not to pass the normality test. This may indicate that the laboratories are more likely to control contamination and other factors that may result in unusual results or multiple distributions that would cause the test to fail.

The majority of analytes/labs that exceeded the target 1% false positive rate by a large amount (i.e., at least a 5% false positive rate) for DL_T based on all blanks, and all analytes/labs that yielded at exceeded the target 1% false positive rate by a large amount for DL_K, had significantly positive skewness, significantly positive kurtosis, or both. For example, Figure 4-5a shows the distribution of calcium results from Laboratory 2, for which the calculated false positive rates for DL_T and DL_K were 13.46% and 8.65%, respectively, based on the full verification set of blank data. These data had levels of skewness and kurtosis that were significantly greater than that which would be expected under a normal distribution.

Milder departures from a normal distribution also yielded higher false positive rates in some cases. For example, Figure 4-5b shows the distribution of vanadium results for Laboratory 2, which yielded a 5.36% rate for DL_T and 3.57% for DL_K.

Distributions that did not significantly deviate from a normal distribution (such as magnesium for Laboratory 3 as shown in Figure 4-5c) tended to produce false positive rates around 1% for DL_T and below 1% for DL_K. Distributions with significantly negative skewness and/or kurtosis (such as Laboratory 2 chromium as shown in Figure 4-5d) tended to have false positive rates below 1% for both DL_T and DL_K.

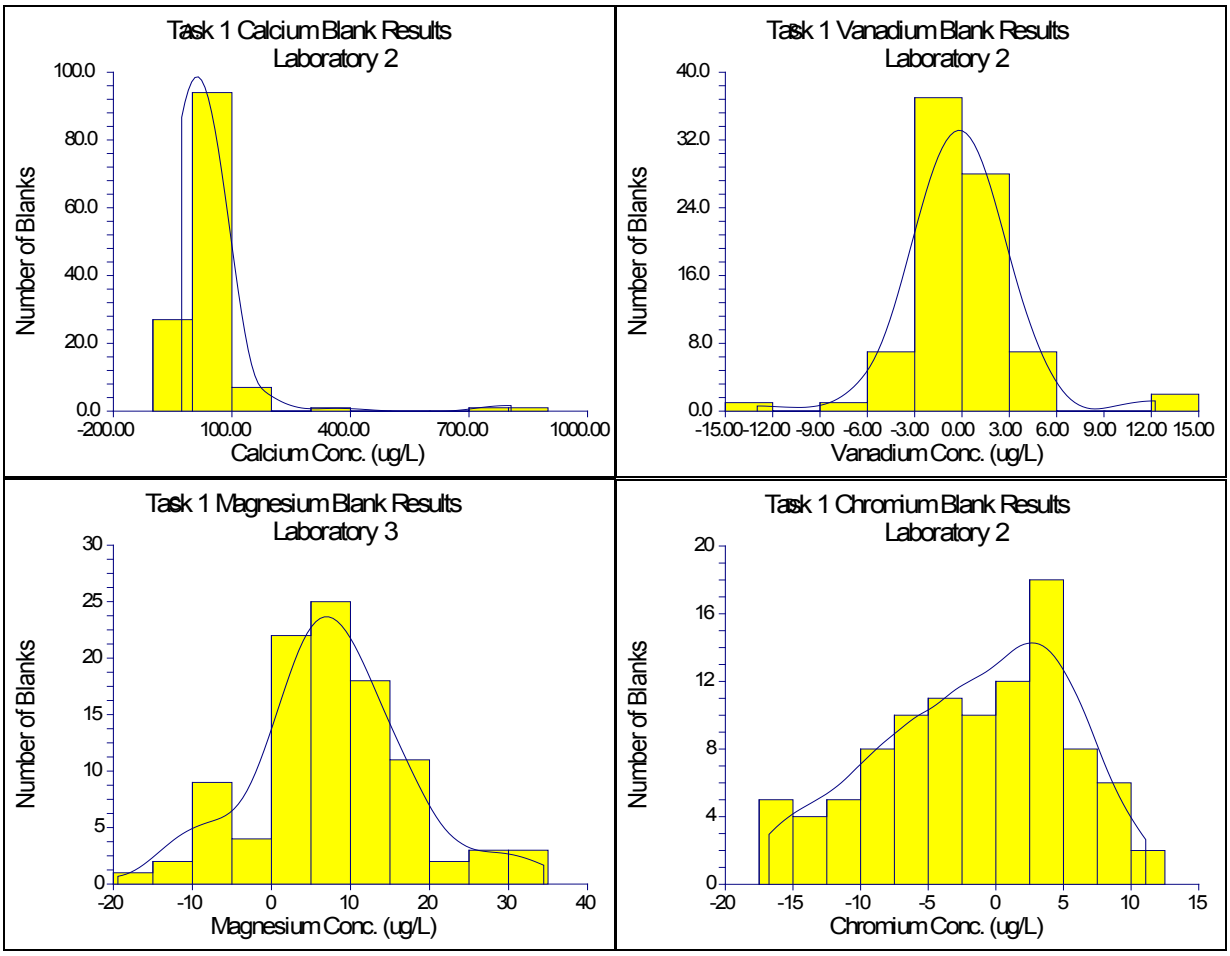

Figure 4-5. Examples of Blank Result Distributions

4.3.2 Outlier Testing

In an EPA-sponsored peer review of the FACDQ 2.4T procedure, multiple reviewers commented on the type and application of the outlier test. In many cases, reviewers suggested alternate outlier tests. Therefore, EPA assessed the effect of these alternate peer-recommended outlier tests on the FACDQ DL calculations. Task 1 blank data were used to perform these assessments, as described below.

The alternate outlier tests included:

- Setting the upper and lower bounds to 3 standard deviations outside the mean.
- Setting the upper and lower bounds to the 97.5th percentile t-statistic outside the mean, with the degrees of freedom set to the number of blanks used in the assessment minus 1.
- Setting the upper and lower bounds to the median ± 4*MAD, where MAD corresponds to the median of the absolute differences between the individual results and the median. This approach is less sensitive to departures from a normal distribution.
- Grubbs test, a frequently used outlier test

EPA initially performed each of these outlier tests on just the Task 1C blank data, applying the FACDQ 2.4T rule that no more than 5% (i.e., 1 out of the 20 Task 1C blanks) may be removed per analyte/laboratory set. The frequency of Task 1C outlier removal based on the different tests is presented in Table 4-6 below:

Table 4-6. Frequency of Task 1C Outlier Removal

Test	# Low Outliers Removed	# High Outliers Removed	# Total Outliers Removed
Mean ± 2SD (FACDQ 2.4T outlier procedure)	18	39	57
Mean ± 3SD	0	10	10
Mean ± t SD	16	35	51
Median ± 4MAD	12	27	39
Grubbs test	3	17	20

As one would expect, there was an inverse relationship between the number of outliers removed and the width of the bounds determined for the specific outlier test. For a set of 20 blanks, the multipliers for the mean ± t SD test and Grubbs test are 2.1 and 2.71, respectively. As a result, the number of analyte/laboratory sets with an outlier removed for the mean ± t SD test was close to that of the mean ± 2SD test, and the number of analyte/laboratory sets with an outlier removed for Grubbs tests was close to that of the mean ± 3SD test. The median ± 4MAD test was the only test that did not always have consistently wider or tighter bounds than the other tests, because the relationship between the MAD and standard deviation was not consistent between sets. This test identified an outlier for 39 of the 71 analyte/lab sets. The effect of the FACDQ 2.4T procedure's 5% maximum removal rule likely mitigated differences in the outlier tests, as more than one result per analyte/lab set was frequently identified as an outlier for the mean ± 2SD and median ± 4MAD tests.

For the tests that identified fewer outliers, the resulting DLs were higher because a result exceeding the Task 1B DL was less likely to be classified as an outlier. As shown in Table 4-7, the mean ± 3SD outlier test (which yielded the fewest outliers) led to an adjustment to DL_T and/or DL_K more frequently than the other outlier tests.

Table 4-7. Frequency of Task 1C DL Adjustment After Application of Various Outlier Tests

Test	% Analyte/Lab Sets With Task 1C DL Adjustment	
	DL_T	DL_K
None	33.8%	18.3%
Mean ± 2SD (FACDQ 2.4T outlier procedure)	21.1%	7.0%
Mean ± 3SD	26.8%	9.9%
Mean ± t SD	21.1%	7.0%
Median ± 4MAD	21.1%	8.5%
Grubbs test	23.9%	9.9%

EPA compared the DLs determined after application of the different outlier tests to the full verification set of blanks to calculate false positive rates; results of this assessment are shown in Table 4-8.

Table 4-8. Estimated False Positive Rates after Application of Various Outlier Tests to Task 1C data

Outlier Test	Overall FP Rate for DL_T	Overall FP Rate for DL_K
None	2.58%	0.97%
Mean ± 2SD (FACDQ 2.4T Outlier procedure)	3.42%	1.13%
Mean ± 3SD	3.14%	1.09%
Mean ± t SD	3.42%	1.13%
Median ± 4MAD	3.32%	1.13%
Grubbs test	3.28%	1.09%

Generally, the rates did not vary greatly between outlier tests for either DL_T or DL_K. The rates exceeded 1% for DL_T and were close to 1% for DL_K, regardless of which outlier test was run. It should be noted that the outlier tests described above were applied to the Task 1C data but not to the limit verification data. To assess how this affected the calculated rates, EPA also applied the various outlier tests to full verification set of blank data.

After excluding the Task 1B and 1C blanks, there were between 51 and 104 blank results remaining per analyte/lab set. Therefore, between 2 and 5 blanks could be removed per set without exceeding the 5% maximum in the FACDQ 2.4T procedure. Unlike the other outlier tests, Grubbs test is run recursively, i.e., the most extreme outlier is removed, the test is re-run, then the next most extreme outlier (if any) is removed, etc. As a result, assessing which tests are more likely to remove outliers is not as simple as comparing the widths of the bounds for the different tests. However, it is worth noting that the 97.5^{th} percentile t-statistic is less than 2 when applied to at least 62 blanks, so this test would have the tightest bounds for most analyte/lab sets in the full verification set.

The rate of outlier removal for each test, and the false positive rate determined after the outliers identified for that test were removed, is presented in Table 4-9.

Table 4-9. Outlier Removal and False Positive Rates for each Outlier Test Applied to Task 1C and the Full Verification Set Data

Test	% Results Identified as Outliers	FP Rate DL_T	FP Rate DL_K
Mean ± 2SD (FACDQ 2.4T Outlier Procedure)	4.3	1.93	0.37
Mean ± 3SD	1.1	2.59	0.60
Mean ± t SD	4.4	1.91	0.37
Median ± 4MAD	3.1	2.05	0.23
Grubbs test	1.3	2.50	0.34

Although the false positive rates are lower than those presented in Table 4-8, they still do not vary widely between outlier tests. This suggests that using a different outlier test from the one used in the FACDQ 2.4T procedure will have only a minor effect on the ability of the procedure to accurately determine limits that meet the target false positive rate.

4.3.3 Temporal Variability

4.3.3.1 Comparison of Blanks over Multiple Study Phases (Tasks)

The FACDQ 2.4T procedure includes a process for verifying DLs and QLs using data analyzed over a period of a year. Due to practical considerations, spiked sample analyses in this study were spread out over a much shorter time period. However, the blank data submitted for Task 1 of this study did cover a longer period of time, as shown in Table 4-10.

Table 4-10. Time Range Covered by Task 1 Blanks

Laboratory	Overall Time Range (days)	Time Range of Task 1B DL Calculation Blanks (days)	Time Range of Task 1C DL Verification Blanks (days)
1	169	20	43
2	148	1	76
3	135	2	25

The amount of temporal variability covered by the seven Task 1B blanks ranged widely between laboratories. The Task 1B time ranges for Laboratories 2 and 3 were more typical of short-term DLs, with the seven blanks analyzed within 1-2 days. For Laboratory 1, the Task 1B time range was much larger (20 days). The false positive rates calculated for Laboratory 1 (1.1% and 0.4% for DL_T and DL_K, respectively, based on the full verification set) were lower than for Laboratory 2 (4.1% and 1.5% for DL_T and DL_K, respectively, based on the full verification set) or 3 (4.3% and 1.1% for DL_T and DL_K, respectively, based on the full verification set), suggesting that the greater amount of temporal variability resulted in higher limits.

The time range for Task 1C was intended to simulate the ongoing verification steps of the FACDQ procedure. The time span for these 20 blanks ranged between 25 and 76 days, and was at least two times as long as the Task 1B time range for each laboratory.

EPA compared the variability between the Task 1B and Task 1C blanks for each analyte/laboratory set using either the F-test (for analyte/laboratory sets that passed the normality test described in Section 4.3.1) or the modified Levene's test (for analyte/laboratory sets that did not pass the normality test described in Section 4.3.1). Given the wider time range covered in

Task 1C, it could be assumed that the variability among those blank data would frequently be significantly greater compared to the variability among the Task 1B blank data. Further, because that time difference is greatest for Laboratory 2, it could be expected that the frequency of significant differences would be greatest for that laboratory's data.

Contrary to expectation, variability did not differ significantly between Task 1 phases for the vast majority of analyte/laboratory sets. A significant difference was observed for only five analyte/laboratory sets (lead, nickel and selenium for Laboratory 1; arsenic and potassium for Laboratory 2). Of these, there was only one case where the variability was significantly greater in Task 1C than in Task 1B (potassium for Laboratory 2); for the other four sets, the variability in Task 1C was significantly lower than in Task 1B. The false positive rates determined for both DL_T and DL_K were above 1% for this analyte/laboratory (6.7% for DL_T and 2.9% for DL_K). Possible explanations for the infrequent significant differences in variability between Tasks 1B and 1C blanks include:

- The small number of blanks used in these tasks resulted in reduced statistical power of the comparisons. With decreased statistical power, a large difference in the variances is needed to yield a significant difference in results; however, a difference in the variances that is not statistically significant may still have a practical effect on the DLs calculated using these data.
- Method 200.7 generally requires fitting new calibrations on a daily basis. As a result, data generated over as little as two days would include more sources of variability than would data from other methods that require daily calibration verification instead of fitting a new calibration.
- The time period covering Task 1B could be unique, such that the short-term variability covered by those seven blank results would not be representative of the variability routinely observed over a time period of that length.

EPA also compared variability between the combined Task 1B and 1C limit calculation data and the full verification set of data using the same variability tests described above. There were six analyte/lab sets for which the full verification set of blanks were significantly more variable than the Task 1B and 1C limit calculation blanks, and four analyte/lab sets for which the full verification set of blanks were significantly less variable than the limit calculation blanks. Laboratory 1, which covered the longest time frame during Task 1B, was the only lab for which no analytes had significantly greater variability in the full verification set. Among the six analyte/lab sets for which the full verification set had significantly greater variability, calculated false positive rates for DL_T tended to exceed 1%, with rates above 5% for four of the sets; however, no false positives were observed for DL_K for any of these six sets. Among the four analyte/lab sets for which the verification set had significantly less variability, calculated false positive rates for DL_T and DL_K tended to be at or below 1%. These results suggest that while significant changes in variability between the data used to calculate the limits and the data subsequently compared to the limits were rare, they may still have affected the overall false positive rates determined for DL_T and DL_K.

4.3.3.2 Comparison of Variability over Different Timeframes

As stated in the previous section, the FACDQ 2.4T procedure requires analyzing blank and spiked samples on an ongoing basis, and recalculating and/or re-evaluating limits at least annually. Therefore, some limits determined using this procedure as currently written could be based on as much as a full year of temporal variability.

Although the existing blank data compiled for this study only covered a maximum of six months, these data can still be used to assess the effect of different analysis timeframes on the resulting variability. The results of this assessment can give an indication of how the compressed analysis timeframe for spiked samples in this study compares to the full timeframe specified in the procedure itself. Additionally, the results of this assessment can answer the question of how the sample analysis and limit evaluation frequencies will impact the calculated limits.

The frequency of blank analyses throughout the time period covered in the historical blank datasets varied for each of the three laboratories. The blank data submitted by each of the laboratories included days for which multiple blanks were analyzed, as well as longer periods of time for which no blanks were analyzed. Therefore, one-day periods could be identified for each laboratory with multiple blank results per analyte. Similarly, longer periods of time also could be identified for each laboratory with multiple results per analyte. For example, there were 23 one-day periods and 12 three-day periods with multiple blank results among the data submitted by Laboratory 1.

To assess the effect of time on analytical variability, EPA selected six different time ranges, as shown in Table 4-11. For each of these ranges, EPA identified a set of non-overlapping time periods for each laboratory with data from multiple blank analyses in each range. Not all periods included analyses on the first and last days of that period; as a result, the actual time ranges used in the assessment were occasionally slightly higher or lower than the target time range. The number of these periods that were identified for each laboratory is shown in the last three columns of Table 4-11. Because not all analytes were analyzed for all blanks, not all periods could be used for all analytes.

Table 4-11. Number of Analytical Time Range Periods Assessed Using Blank Data

Time Range	Actual Time Ranges Used in Assessment	Maximum Number of Discrete Periods per Laboratory *		
		Lab 1	Lab 2	Lab 3
1 day	1 day	23	31	30
3 days	2-3 days	12	12	12
7 days	6-8 days	10	9	11
14 days	12-15 days	7	5	7
28 days	25-32 days	3	4	3
6 months	135-169 days	1	1	1

* Not all periods useable for all analytes due to missing or invalid data

EPA calculated the variance of the blank results for each analyte for each discrete period. After these variances were calculated, pooled estimates were calculated for each analyte, laboratory, and time range. For example, the 23 one-day aluminum variances calculated from Laboratory 1's data were pooled into a single variance estimate, and the 12 three-day aluminum variances

calculated from Laboratory 1's data were pooled in to a second variance estimate. EPA compared the pooled estimates of temporal variability to each other using either the F-test or modified Levene's test, depending on whether that data had passed the normality test in Section 4.3.1. Results of the variability comparisons between time ranges are presented in Table 4-12.

Table 4-12. Percentage of Analytes/Lab Sets with Significantly Greater Variability in Longer Time Range

	1 day	3 days	7 days	14 days	28 days	6 months
1 day	X	38%	70%	73%	75%	86%
3 days		X	8%	15%	27%	39%
7 days			X	3%	8%	18%
14 days				X	3%	6%
28 days					X	3%
6 months						X

As would be expected, significant increases in variability were observed most frequently when comparing longer time ranges to the shortest time ranges. For example, analyses spread out over approximately one week were significantly more variable than analyses spread out over one day for 70% of the analyte/laboratory sets. Almost every analyte/laboratory set (86%) yielded significantly greater variability in the full six-month time range than in the one-day range. However, only a few sets yielded significantly greater variability in the full time range compared to the 7-day (18%), 14-day (6%) and 28-day (3%) ranges. This supports the study design assumption that compressing Task 2 and 3 analyses to two weeks would have little impact on the results.

Figures 4-6 through 4-8 show the pooled standard deviations for six of the analytes (antimony, arsenic, cadmium, copper, lead, and selenium) for each of the three laboratories. For Laboratory 1, variability tends to level off at around seven days for most of the analytes. Variability for all analytes other than antimony significantly increased after one day, but only arsenic and lead showed significant increases in variability after 3 days, and only lead exhibit significant increases after 14 days. Slightly later increases in variability were observed among these analytes for Laboratory 2, with significant increases after seven days occurring for antimony and copper. Similarly, significant increases were observed after 7 days for only arsenic and selenium for Laboratory 3. For all three laboratories, the magnitude of the increase in variability tended to be greatest for most analytes between one day and three days.

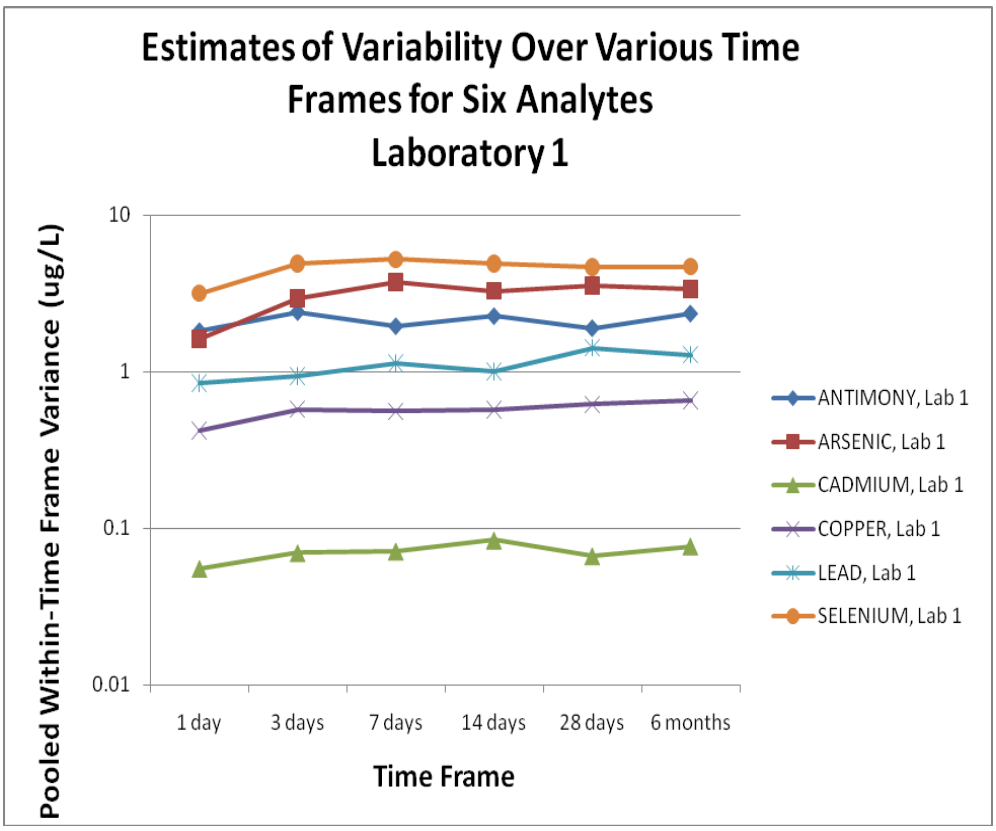

Figure 4-6. Pooled Standard Deviations for Various Time Ranges for Six Analytes – Laboratory 1

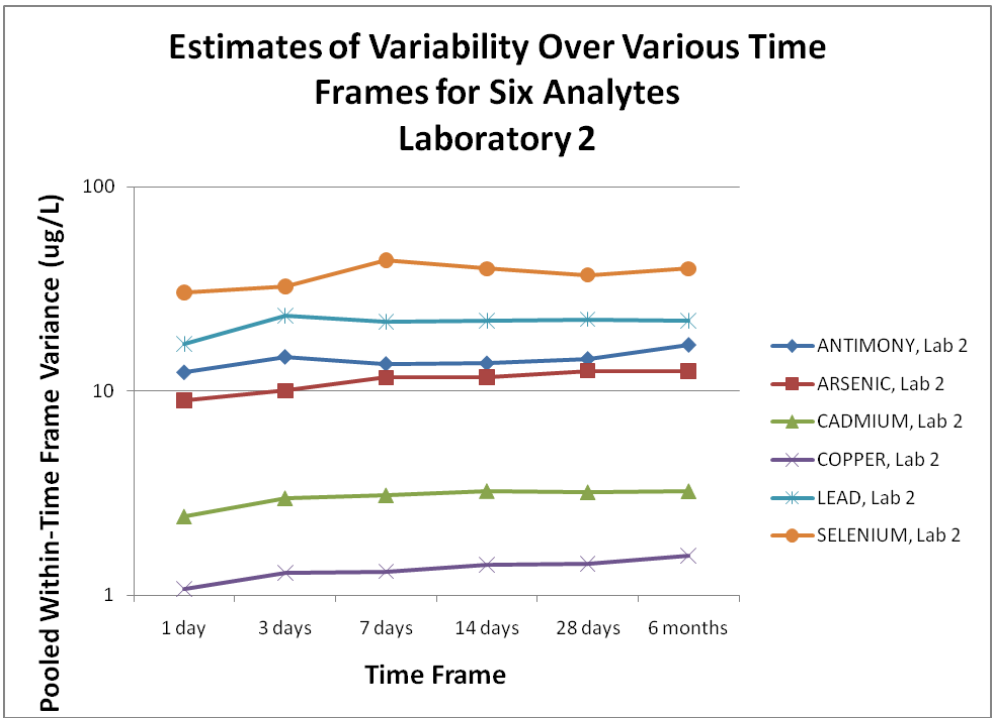

Figure 4-7. Pooled Standard Deviations for Various Time Ranges for Six Analytes – Laboratory 2

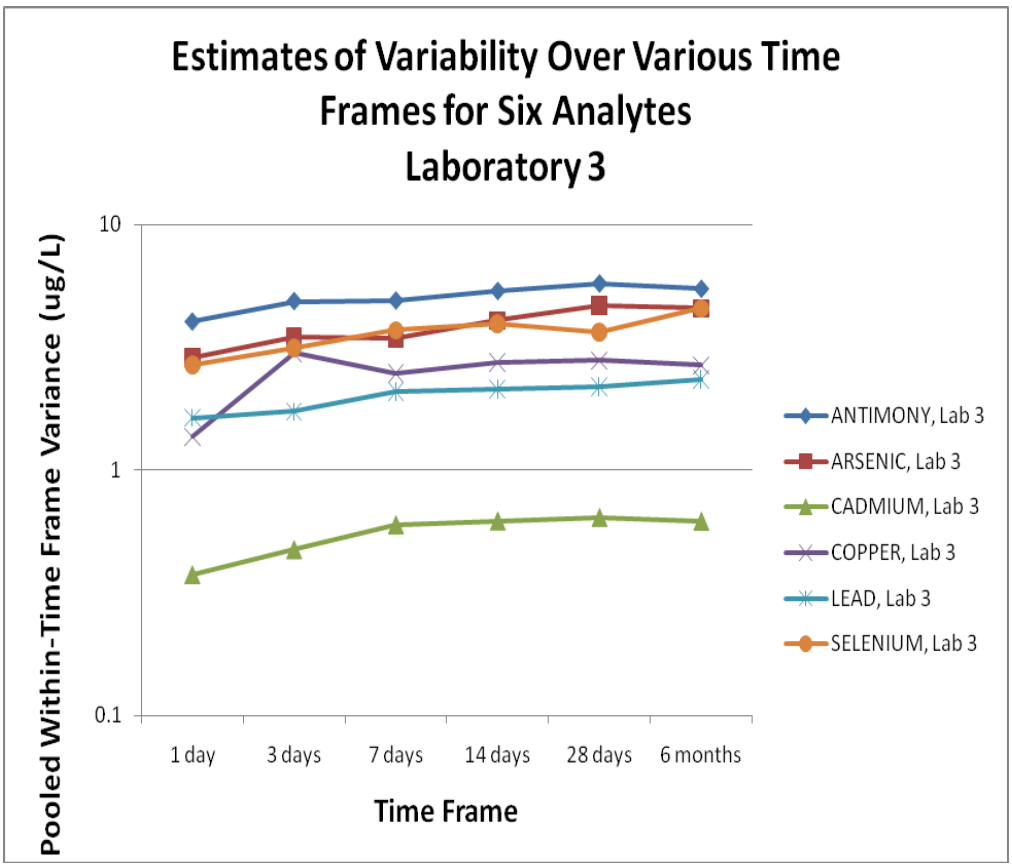

Figure 4-8. Pooled Standard Deviations for Various Time Ranges for Six Analytes – Laboratory 3

4.3.4 Alternate Calculations

Although the DL_T calculation in the FACDQ 2.4T procedure is considered to be an approximate 99% upper prediction limit for a single sample analysis, one peer reviewer noted that the calculation of this limit does not match the exact formula for a prediction limit. The exact calculation for a DL representing an upper 99% prediction limit would be:

$$DL_{PI} = Max(0, \overline{X}) + t_{0.99;n-1)} * \sqrt{(1+\frac{1}{n})} * s$$

Where $\overline{X}$ *is the mean blank,*
s is the standard deviation of the blanks, and
n is the number of blank results used in the calculation.

When calculated using 7 blank results DL_{PI} would be approximately 7% greater than DL_T when the mean blank value is zero or negative, and would be less than 7% greater than DL_T when the mean blank value is positive.

Although the FACDQ DL_T calculation is simpler than the exact prediction limit formula, the slight decrease in the resulting limit could have an effect on the ability to estimate the level at which the 1% false positive rate is met. As a result, EPA calculated and adjusted DL_{PI} using the same data and process that was used for DL_T.

On average, DL_{PI} was 5.07% higher than the corresponding DL_T. The two limits were equal for 14 analyte/laboratory sets; these sets correspond to 14 of the 15 sets for which DL_T was adjusted based on the Task 1C maximum.

EPA assessed the false positive rate for DL_{PI} using both the full set of blanks not used in Tasks 1B or 1C (i.e., the full verification set), and compared the false positive rate to the target 1% using the same bootstrap estimation approach described in Section 4.2. The results of these assessments are presented in Table 4-13.

Table 4-13. False Positive Rates for DL_T calculated using Exact Prediction Interval Formula

Outliers Removed in Verification Data	Total # Blanks	False Positive Rate Based on Full Verification Set	# Total Blanks per Bootstrap Run	Mean False Positive Rate	p-value for Proportion test
No	5,032	3.02	1,420	2.73	<0.001
Yes	4,845	1.67		1.49	0.032

Using the exact prediction limit formula decreased the overall false positive rate based on the full verification set of blanks from 3.42% to 3.02% before outlier removal, and from 2.06% to 1.67% after outlier removal. The false positive rates based on the bootstrap estimation also decreased (from 3.09% to 2.73% before outlier removal, and from 1.86% to 1.49% after outlier removal); however, these limits still were statistically significantly greater than 1% based on the proportion test.

Section 5: Task 2 Limit Calculations

This section describes the laboratory analyses and calculations performed as part of Task 2 of the study; specifically the determination of FACDQ QLs, LCMRLs, and MRLs for each analyte and laboratory. Both the participant laboratories and EPA personnel performed calculations of these limits. Section 5.1 describes the calculation of LCMRLs during the study, Section 5.2 describes the calculation of FACDQ QLs, and Section 5.3 describes the calculation of MRLs. Section 5.4 describes deviations from the temporal variability requirements of the study that occurred during Task 2, and any effect they may have had on the calculated limits.

5.1 LCMRL Determination

5.1.1 Method 200.7

Laboratories were instructed to spike and analyze four replicate samples at seven different concentrations in order to determine an LCMRL for each of the 24 analytes studied. However, Laboratory 2 analyzed 7 replicate samples at 11 different concentrations, and Laboratory 3 analyzed 4 replicates samples at 9 different concentrations. To ensure comparability among laboratories, EPA recalculated LCMRLs for these laboratories using only a subset of their data. Laboratory 2 LCMRLs were recalculated using only the seven lowest spike levels per analyte, and using only the 1st, 3rd, 5th and 7th replicates analyzed at each spike level. Laboratory 3 LCMRLs were recalculated using only the seven lowest spike levels per analyte. If no limit could be achieved using these seven spike levels, the remaining data were treated as "respike" concentrations for the purpose of the assessment, with the lowest remaining spike level added to the calculation one-at-a-time until an LCMRL could be determined. For analytes for which an LCMRL still could not be produced after all additional spike levels had been added to the calculation, the laboratories were instructed to spike and analyze four replicate samples at an additional level. Laboratory 1 (which originally performed the seven spike levels/four replicate design), also spiked and analyzed an additional four replicates at a new concentration for any analytes that did not produce an LCMRL using the original seven spike levels.

Table 5-1 gives the frequency for which the minimum seven spike level/four replicate design could produce an LCMRL for each laboratory performing Method 200.7. An LCMRL could be determined approximately 73% of the time using this minimum design. Among those analyte/laboratory sets for which an LCMRL could not be determined, most required spiking at a lower level (i.e., recoveries across the initial spike range were too precise and accurate to estimate an LCMRL) rather than requiring spiking at a higher level (i.e., recoveries across the initial spike range were not precise and/or accurate enough to estimate an LCMRL).

Table 5-1. Initial Design Success Rate for Determining LCMRLs, Method 200.7

Laboratory	% Analytes with LCMRL determined based on 7 spike levels	% Analytes Needing Higher Spike Level	% Analytes Needing Lower Spike Level	# Analytes for which no LCMRL could be calculated after respiking
1	50	4.2	45.8	1
2	87.5	12.5	0	0
3	82.6	4.3	13.0	0
All	73.2	7.0	19.7	1

After all respiking was performed, an LCMRL could be determined for 70 of the 71 analyte/lab sets. Lower spiking was still required for one analyte (beryllium) for Laboratory 1 to determine an LCMRL; however this was not done due to practical study considerations (laboratories were not required to make a third attempt to determine any of the of limits in this study).

5.1.2 Method 625

The LCMRL procedure does not differentiate between types of analytical methods used to determine an LCMRL. Therefore, laboratories performing Method 625 in the study followed the same steps as laboratories performing Method 200.7. Specifically, the labs were instructed to spike and analyze four replicate samples at seven different concentrations to determine an LCMRL for each of the 49 analytes studied, and for any analytes that did not produce an LCMRL on the initial attempt, the labs were required make a second attempt by spiking and analyzing another four replicates at a new spike level. However, Laboratory 4 spiked four replicate samples at 13 different concentrations on their initial attempt, and Laboratory 5 spiked 4 replicate samples at 10 different concentrations on their initial attempt. EPA recalculated LCMRLs for these two laboratories using only the seven lowest spike levels per analyte. If no limit could be achieved using these seven spike levels, the remaining data were treated as "respike" concentrations for the purpose of the assessment, with the lowest remaining spike level added to the calculation one-at-a-time until an LCMRL could be determined. Table 5-2 gives the frequency for which the minimum 7 spike level/ 4 replicate design could produce an LCMRL for each laboratory for Method 200.7.

Table 5-2. Initial Design Success Rate for Determining LCMRLs, Method 625

Laboratory	% Analytes with LCMRL determined based on 7 spike levels	% Analytes Needing Higher Spike Level	% Analytes Needing Lower Spike Level	# Analytes for which no LCMRL could be calculated after respiking
4	30.6	69.4	0	24
5	57.1	36.7	6.1	4
6	8.2	0	91.8	0
All	32.0	35.4	32.7	28

Across the three laboratories, an LCMRL could be determined based on the minimum 7 spike level/4 replicate design approximately one-third of the time (47 of the 147 analyte/lab sets), at least one higher spike level was needed approximately one-third of the time, and at least one lower spike level was necessary one-third of the time. After respiking was completed, as many as 14 spike levels per analyte were available for the LCMRL determination, depending on the laboratory. As stated above, when more than one additional spike level was available, the lowest

remaining spike level was added one at a time, until a limit could be produced, or no remaining spike levels were available.

Among the 100 analyte/lab sets for which an LCMRL could not be determined with the initial seven spike levels, an LCMRL could be determined using some or all of the additional spike levels for 82 sets. In the majority of the remaining 18 cases, the range of spike levels analyzed encompassed the laboratory's full calibration range, indicating that the LCMRL MQO (50-150% individual sample recovery with 99% probability) was not achievable for that analyte for that laboratory. An example of this is shown in Figure 5-1 (3,3'-Dichlorobenzidine for Laboratory 4); this figure shows the recovery plot produced by the EPA's LCMRL calculator for this analyte. Sample recoveries for this analyte, and the calculated lower prediction limit based on those recoveries, tended to be well below 50% across all concentrations. In other cases, the recovery and precision of the data seemed to indicate that the MQO was achievable even though no limit was produced by the software.

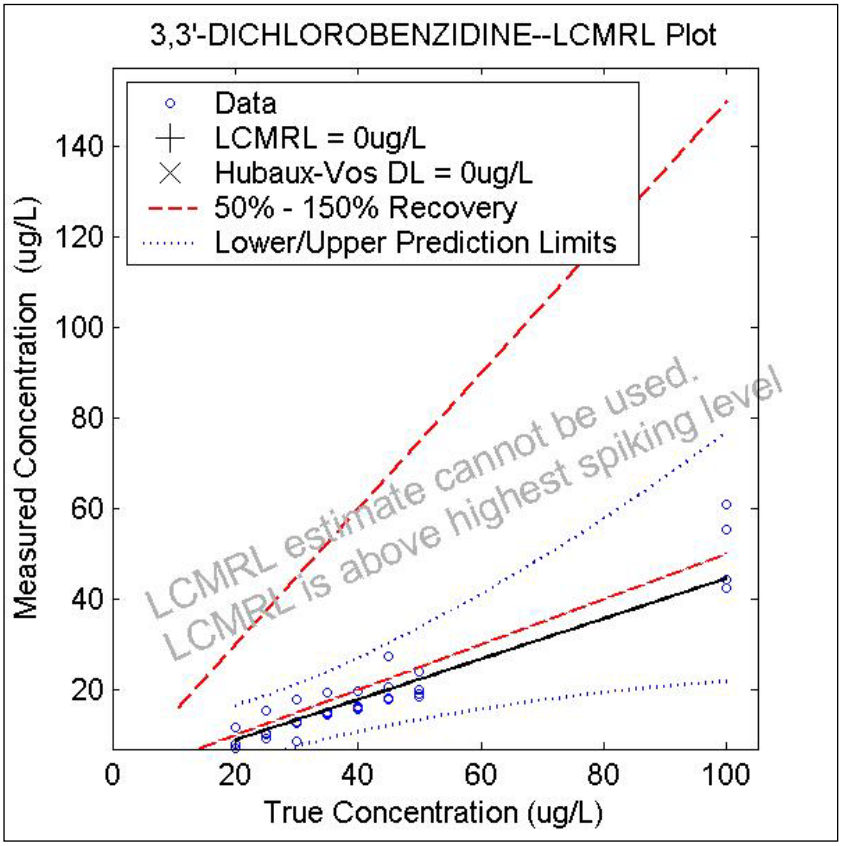

Figure 5-1. EPA LCMRL Calculator Recovery Plot – Lab 4 3,3'-Dichlorobenzidine

5.1.3 Recalculation of LCMRLs

As noted above, the LCMRL calculator output for some analyte/lab sets that could not produce an LCMRL appeared to be at odds with the method performance observed in the data. Upon investigation of these sets, EPA determined that the issue was due to a minor change to the LCMRL calculation algorithm (a modification of the Tukey biweight factor criterion used to

assess model convergence). Though the study was initially designed to reflect the LCMRL calculation procedure available in EPA's downloadable LCMRL software, EPA assessed the LCMRL procedure after the Tukey biweight factor was switched back to its original value because EPA had decided that this algorithm was to be used in future LCMRL calculations. Therefore, EPA recalculated all LCMRLs using the revised algorithm (i.e., the version not used by the LCMRL software), and compared these revised limits to the limits originally determined using the software algorithm.

In most cases, the effect of the algorithm modification on the resulting LCMRL was slight. The median RPD between the original and recalculated LCMRLs for Method 625 was 3.4%, and the RPD exceeded 20% for only 13 of the 115 sets for which an LCMRL could be determined using both algorithms. For Method 200.7, the median RPD between the original and recalculated LCMRLs was 4.9%, and the RPD exceeded 20% for only 7 of the 68 sets for which an LCMRL could be determined using both algorithms. Although the revised algorithm was able to yield more LCMRL values for Method 200.7, it yielded fewer for Method 625. The revised algorithm was able to produce an LCMRL for two of the three Method 200.7 sets that did not yield an LCMRL using the original software algorithm. The revised algorithm also was able to produce an LCMRL for 5 of the Method 625 sets that did not produce an LCMRL using the original software. However, the original software was able to produce a Method 625 LCMRL for 14 data sets that did not yield an LCMRL using the revised algorithm.

5.2 FACDQ QL Determination

Unlike the LCMRL procedure, the FACDQ procedure differentiates between types of analytical methods (i.e., uncensored methods such as Method 200.7 and censored methods such as Method 625). Although the procedures for determining the QLs are similar for both types of methods, the procedures differ in their approaches to determining the DLs. These differences led to determination of two different DL values (DL_t and DL_k) for Method 200.7 during Task 1 as described in Section 4. Two corresponding QL values (QL_{DLT} and QL_{DLK}) also can be calculated for these limits, as described in Section 5.2.1 below. For censored methods, the QLs are determined before the DLs, and both the 2.4 and 2.4T versions of the FACDQ procedure rely on a single version of each limit. Section 5.2.2 describes how the Method 625 QLs were determined; a description of how the Method 625 DLs were calculated is provided in Section 6. The FACDQ procedure specifies use of downspiking to select spike levels for determining QLs; a discussion of this downspiking procedure is provided in Section 5.2.3.

5.2.1 Method 200.7

Laboratories were instructed to spike and analyze seven replicate samples at a single concentration to determine the FACDQ QL_{DLT} for each analyte, and at a single concentration to determine the FACDQ QL_{DLK} for each analyte. Laboratories also were instructed to select these spike levels based on the results of downspiking analyses as described in Section 2.4.1 of the FACDQ 2.4T procedure. Among the three laboratories performing Method 200.7, Laboratory 1 spiked at two different levels per analyte, Laboratory 2 spiked at a single spike level per analyte, and Laboratory 3 spiked at 4-8 levels per analyte. How these spike levels were chosen, and the effect of any discrepancies from the downspiking instructions in the procedure, are discussed in Section 5.2.3.

Downspiking Requirements. EPA compared the spike levels chosen by Laboratories 1 and 2 to their downspiking data and FACDQ DLs to determine which spike levels were appropriate for which QL. In order to meet the FACDQ requirements, the spike levels selected must be at least 2x greater than the corresponding DLs determined in Task 1.

- For Laboratory 1, EPA determined that:
  - The lower spike level was appropriate for both QLs for 7 analytes,
  - The lower spike level was appropriate for QL_{DLT} and the upper level was appropriate for QL_{DLK} for 9 analytes,
  - The higher spike level was appropriate for both QLs for 5 analytes, and
  - For three analytes, neither of the spike levels was appropriate, and further spiking was necessary for at least one of the QLs.
- For Laboratory 2, EPA determined that the chosen spike level was appropriate for QL_{DLT} for 19 of the 24 analytes, but was appropriate for QL_{DLK} for only 11 of the 24 analytes. As a result, further spiking was necessary for Laboratory 2 for 19 analytes.

Laboratory 3 did not provide downspiking data (see Section 5.2.3). Instead, Laboratory 3 spiked each analyte at either 4 or 8 levels, depending on whether DL_T=DL_K. When the two DLs were not equal, Laboratory 3 spiked at a total of 8 levels. When the two DLs were equal, Laboratory 3 spiked at a total of four different levels, but spiked separate sets of seven replicates for QL_{DLT} and QL_{DLK}. The laboratory spiked at exactly 2x the corresponding DL, and then at factors of 3 and 9 above this level (i.e., 6x and 18x the DL), and at a factor of 3 below this level (i.e., 2/3x the DL). In order to evaluate the FACDQ QLs for this laboratory, EPA used the 2x DL spike levels as the initial spike level.

MQO Requirements. In addition to meeting the downspiking requirements (including the requirement that the QL spike level be at least 2x the DL), each laboratory's spiked sample results also had to meet both MQO requirements (20% RSD and 70-130% recovery) in order to produce valid QLs. Three analytes from Laboratory 2 and four analytes from Laboratory 3 failed to meet at least one of the MQOs for QL_{DLT}. Specifically, the RSD exceeded 20% for six analytes, and the mean recovery fell below 70% for one analyte.
In practice, Laboratory 3 did not need to respike because the laboratory had already analyzed at higher levels during their initial round of spiking. Instead, EPA used the next higher spike level originally submitted (i.e., the level at 6x the DL) as the respiking level for the purpose of the QL assessment. Laboratory 2 spiked and analyzed 21 replicate samples at the new spike level, rather than the 7 that were required. For the purpose of assessing whether the MQOs were met by Laboratory 2 during respiking, EPA randomly selected 7 of the 21 replicates, while also ensuring that at least one replicate from each preparation batch was selected to maximize temporal variability.

Overall Achievement of MQOs after Respiking. The respiking levels chosen by the laboratories met the 2x DL requirements for all but four analytes (aluminum, barium, magnesium, and lead for Laboratory 2). The mean recovery and RSD MQOs were met for all respiked analytes for all labs. Table 5-3 shows descriptive statistics of the mean biases and RSDs at the final spike level for each analyte/laboratory set (i.e., the initial spike level if it met the MQOs and downspiking requirements, or the respiked level if any requirements were not met initially). Mean bias was

calculated as the absolute difference between mean recovery and 100%; EPA used this value in the assessment rather than mean recoveries to avoid the possibility of mean recoveries below 100% and mean recoveries above 100% cancelling each other out. Box plots of the mean biases for QL_{DLT} and QL_{DLK} also are shown in Figure 5-2, and box plots of the RSDs for QL_{DLT} and QL_{DLK} are shown in Figure 5-3.

Table 5-3. Descriptive Statistics of FACDQ QL MQOs – Method 200.7

MQO Statistic	Limit	Number of Analyte/lab sets	Mean	Median	Minimum	Maximum
Mean Bias (%)	QL_{DLT}	71	6.32	5.31	0.07	20.29
	QL_{DLK}	71	5.81	4.13	0.18	23.25
RSD (%)	QL_{DLT}	71	8.57	8.05	1.61	19.65
	QL_{DLK}	71	7.08	5.46	1.00	19.45

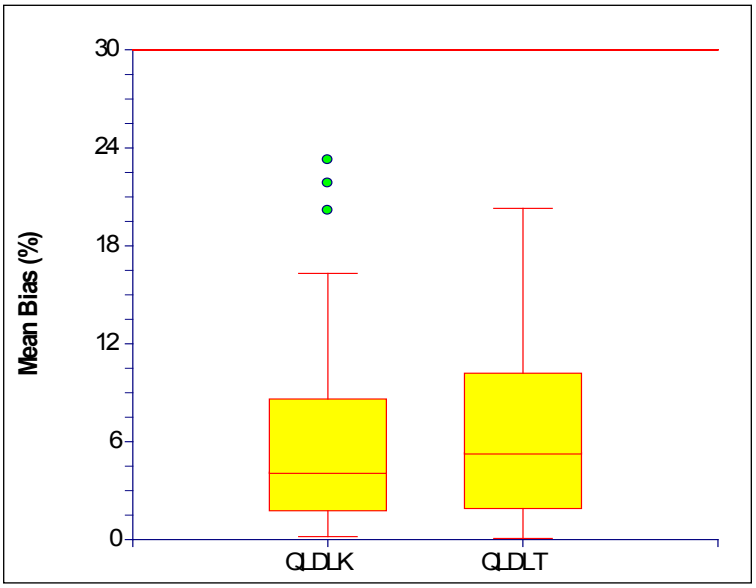

Figure 5-2. Mean Bias at Level where MQOs Passed - QL_{DLT} and QL_{DLK}

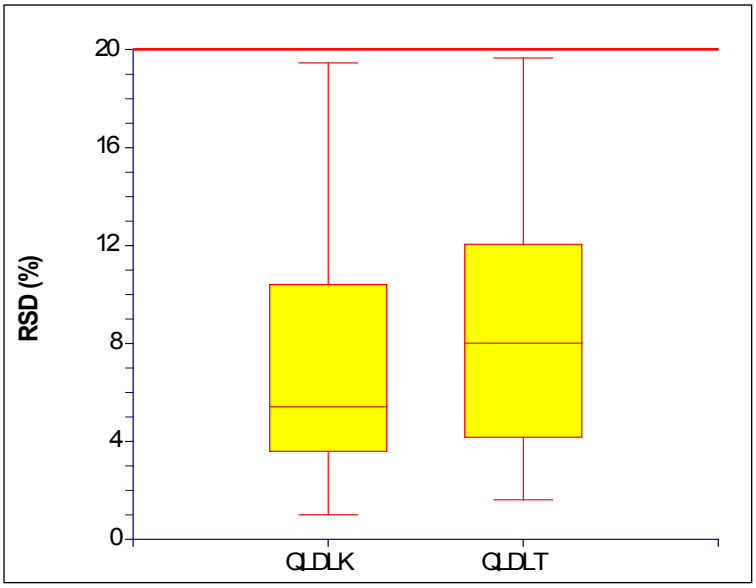

Figure 5-3. RSDs at Level where MQOs Passed - QL_{DLT} and QL_{DLK}

Because only data that met MQOs were included in the above table, the mean and median of the RSDs and mean biases were well below the cutoffs of 20% and 30%, respectively. Generally, the mean and median of the MQO statistics were slightly lower for the QL_{DLK} than for the QL_{DLT}.

Once the spike level had been identified at which the MQOs all were met, the lowest expected result (LER) assessment was performed according to FACDQ 2.4T instructions. Specifically, the LER was calculated for each analyte and compared to the corresponding FACDQ DL. For all analyte/lab sets, the LER calculated at the final QL_{DLK} and QL_{DLT} spike levels exceeded the corresponding DLs, and no further QL adjustments were necessary.

5.2.2 Method 625

As with the FACDQ QL assessment for Method 200.7, laboratories were instructed to initially spike and analyze seven replicate samples at a single concentration when determining the FACDQ QL, with the spike level chosen based on the results of downspiking analyses as described in Section 3.2.1 of the FACDQ 2.4T procedure. Among the three laboratories performing Method 625, Laboratories 4 and 5 spiked and analyzed every analyte at three different levels, and Laboratory 6 spiked every analyte at a single level. How these spike levels were chosen, and the effect of any discrepancies from the downspiking instructions in the procedure, are discussed in Section 5.2.3.

Downspiking Requirements. EPA compared the spike levels chosen by Laboratories 4 and 5 to their downspiking data and the FACDQ 2.4T procedure requirements to determine the appropriate initial QL level for each analyte. Ideally, this spike level would be the lowest spike level that was at least two times the highest blank result identified during Task 1 that was also within the calibration range of the instrument and yielded an instrument signal that met qualitative identification during the downspiking analyses. There was only one analyte/lab set (2,4-Dinitrophenol for Laboratory 4) for which none of the submitted spike levels met the

procedure downspiking requirements; specifically, all spike levels for this analyte were below the lowest calibration standard included in the calibration linearity assessment for this analyte. As a result, Laboratory 4 performed further spiking for 2,4-Dinitrophenol.

MQO Requirements. For all other analyte/lab sets, the spike level that was chosen based on the downspiking requirements in the FACDQ procedure was identified as the initial spike level, and the mean recoveries and RSDs were calculated at this level for each analyte and compared to the target MQOs (30% RSD and 40-160% recovery). A total of 16 analyte/lab sets (or 10.9% of the total number of sets), including ten analytes for Lab 4, four analytes for Lab 5, and two analytes for Lab 6, failed to meet the target MQOs at this starting level. In all 16 cases, the calculated RSD exceeded 30%; the mean recovery fell outside the 40-160% bounds for 9 of the 16 sets.

Laboratories 4 and 5 had both spiked at multiple levels during their initial round. Therefore, results of the remaining initial spike levels were used to assess the analytes that failed MQOs. For Laboratory 4, MQOs for five analytes still were not met with either of the remaining spike levels, and for Laboratory 5, MQOs for two analytes still were not met. Laboratory 6 had initially spiked at only one level (as required), therefore, this laboratory needed to re-spike at a new level for their 2 analytes that failed to meet MQO requirements. However, due to a data entry error in the calculations file, the failed MQO for one of these two analytes was not identified until the respiking phase already had been completed, and respiking for this analyte could not be performed. Additionally, one analyte failed to meet the 30% RSD MQO after some of the data were subsequently reassessed using data from a greater temporal timeframe (See Section 5.4). As a result, additional spiking was performed for a total of 8 lab/analyte sets.

Overall Achievement of MQOs after Respiking. Each laboratory spiked seven replicate samples at one or two additional spike levels for their analytes that did not meet MQOs using any of the data available from the first round of spiking, and the MQOs were achieved at one or both of the levels for all 8 of these respiked analyte/lab sets. Table 5-4 shows the descriptive statistics of the mean biases and RSDs at the final spike level for each analyte/laboratory set (i.e., the initial spike level if it met the MQOs and downspiking requirements, or the respiked level if any requirements were not met initially). Mean bias was calculated as the absolute value difference between mean recovery and 100%; EPA used this value in the assessment rather than mean recoveries to avoid the possibility of mean recoveries below 100% and mean recoveries above 100% cancelling each other out. The distribution of the mean bias and RSD at the FACDQ QL are shown in Figures 5-4 and 5-5, respectively.

Table 5-4. Descriptive Statistics of FACDQ QL MQOs – Method 625

MQO Statistic	Number of Analyte/lab Sets	Mean	Median	Minimum	Maximum
Mean Bias (%)	144	22.2	19.9	0.42	59.4
RSD (%)	144	13.1	13.2	3.60	29.6

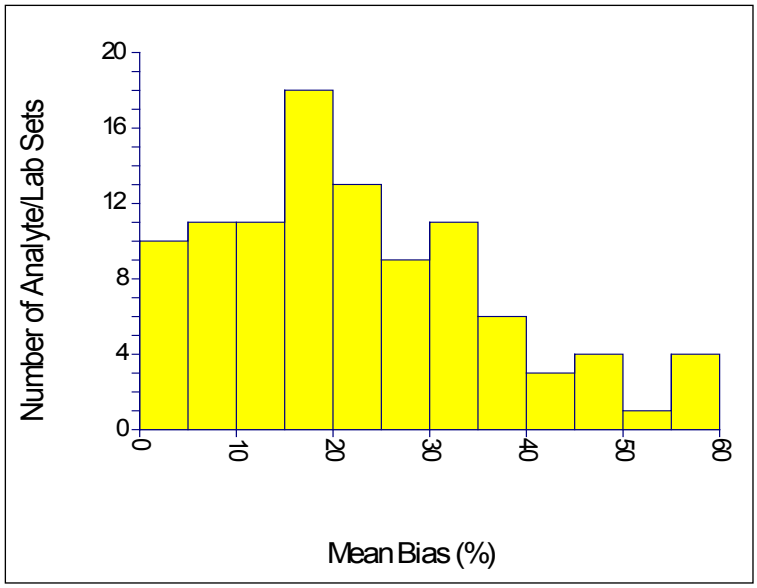

Figure 5-4. Distribution of Task 2 Mean Bias - Method 625

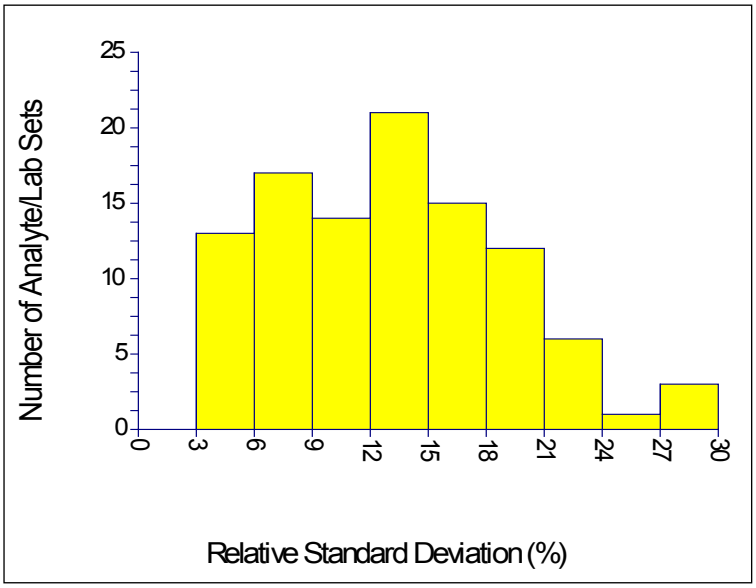

Figure 5-5. Distribution of Task 2 RSD - Method 625

Adjustment of QLs based on LERs. Data from the spiked sample results for Method 625 also were used to calculate the FACDQ DL. Further discussion on this calculation is presented in Section 6, DL Assessment for Method 625. These DLs then were compared to the LER calculated at the final QL spike level, as instructed by the FACDQ 2.4T procedure. The DL exceeded the calculated LER for a total of 18 lab/analyte sets, including 8 analytes for Laboratory 4, 4 analytes for Laboratory 5, and 6 analytes for Laboratory 6. For most of these sets, the MQOs were achieved at the initial spike level, and therefore the DL and LER were calculated using the same set of seven replicates. When this occurs, the resulting DL exceeds the LER when the RSD is greater than or equal to approximately 20%. For each of these 18 sets, the QL was then adjusted based on the calculated DL. The resulting increase in the QL ranged between 1.8% and 51.1%.

5.2.3 Assessment of FACDQ QL Downspiking Analyses

In Task 2 of the study, laboratories were instructed to follow the steps described in Section 2.4.1 (Method 200.7) or Section 3.2.1 (Method 625) of the FACDQ 2.4T procedure when choosing a starting spike level for the QL(s). Specifically, Section 2.4.1 of the procedure states that for Method 200.7, the laboratories must do the following:

> *Using the laboratory's knowledge of the method, analyze spikes of the analyte(s) in blanks. Start at a measurable concentration and reduce the spike concentrations successively in steps of approximately 3 (e.g., 100, 30, 10, 3, 1 etc) until:*
>
> - *Signal to noise ratio is less than 3, or*
> - *Qualitative identification criteria are lost, or*
> - *Signal is lost, or*
> - *The spike concentration is less than twice the detection limit determined in Section 2.2-2.3.*
>
> *Use the lowest spiking concentration, at or above the lowest calibration standard, at which none of the above occur. The chosen spike level is the current QL estimate.*

Similarly, Section 3.2.1 of the procedure states that for Method 625, the laboratories must do the following:

> *Using the laboratory's knowledge of the method, analyze spikes of the analyte(s) in blanks. Start at a measurable concentration and reduce the spike concentrations successively in steps of approximately 3 (e.g., 100, 30, 10, 3, 1 etc) until:*
>
> - *Signal to noise ratio is less than 3, or*
> - *Qualitative identification criteria are lost, or*
> - *Signal is lost, or*
> - *The value is approximately 2x the highest blank result observed for blanks that yielded a signal in Section 1.*
>
> *Use the lowest spiking concentration, at or above the lowest calibration standard, at which none of the above occur. The chosen spike level is the current QL estimate.*

The laboratories were required to analyze these FACDQ downspiking samples and report them with the Task 2 data. However, not all laboratories selected spike levels based on the above criteria. Some laboratories chose initial FACDQ QL spike levels that fell below 2x the corresponding detection limit or Task 1C highest blank, while other laboratories chose initial spike levels that were greater than the lowest concentration at which none of the bulleted items occurred.

Each laboratory was subsequently contacted to determine whether this was due to misinterpretation of the instructions, practical issues such as scheduling, or other issues. Discussions of how the downspiking samples were or were not used are presented in the following sections.

5.2.3.1 Downspiking Assessment - Method 200.7

EPA calculated ratios to compare the initially chosen spike levels to the two Task 1 DLs for each Method 200.7 laboratory and analyte; descriptive statistics of these ratios are presented in Table 5-5.

Table 5-5. Initial Spike Level/DL Ratios Method 200.7

Laboratory	Limit	Geometric Mean Ratio of all Analytes	Minimum Ratio of all Analytes	Maximum Ratio of all Analytes
1	DL_T	1.93	0.03	9.64
	DL_K	1.15	0.03	5.32
2	DL_T	3.28	0.88	9.45
	DL_K	1.81	0.48	5.13
3 [1]	DL_T	2.12	1.94	3.39
	DL_K	2.17	1.92	3.58

[1] Laboratory 3 spiked at 2x the DLs that were provided to them (within rounding) for all analytes. However, the provided DLs were incorrect for a few analytes, and as a result, the ratio was not exactly 2 for all analytes.

Laboratory 3 did not report downspiking results for any analytes. When asked about this, the laboratory stated that they had misunderstood the instructions, and had assumed that seven replicates needed to be prepared for all analytes at multiple levels varying by a factor of 3. Based on this interpretation, the laboratory analyzed seven replicates at exactly 2x the associated DL_T and DL_K that EPA provided to them based on their Task 1 data, as well as at levels above and below this amount. For some analytes, the provided DLs were not the correct values due to calculation errors, so the ratios presented in the above table were not equal to 2 for all analytes; however, the difference in DLs generally did not affect the appropriateness of the chosen spike level. As a result, EPA was able to identify an initial spike level that closely matched the lowest level at which the downspiking criteria could be met.

Laboratories 1 and 2 reported downspiking results along with their Task 2B data, as required. However, the relationship between the downspiking results and the initial QL spike level that was chosen for each analyte frequently was unclear. For Laboratory 1, the initial QL spike level was less than 2x DL_T for 8 analytes and was less than 2x DL_K for 17 analytes. Similarly, the initial QL spike level chosen by Laboratory 2 was less than 2x DL_T for 5 analytes and was less than 2x DL_K for 13 analytes. This contradicts the instructions in Section 2.4.1 of the FACDQ 2.4T procedure. (According to the procedure, higher spike levels should have been chosen for these analytes.) The laboratories performed the necessary analyses at the higher level during the respiking phase. When asked why the inappropriately low spike levels were chosen, the laboratories replied as follows:

- One laboratory stated that they based their downspiking levels on what would normally be used in MDL studies, and as a result, all of their downspiking sample concentrations were too low to meet the 2x DL criterion for some analytes.

- One laboratory stated that the initial spike levels used were the result of a laboratory miscommunication, and samples were spiked at the originally chosen levels (i.e., based on the downspiking analyses) during the respiking phase.

Although Section 2.4.1 of the FACDQ procedure does not explicitly define how far above the previously determined DL the initial QL may be, the intent of the downspiking procedure is to identify the lowest level that is at least twice the DL (and meets the other 3 criteria listed in that section). In some cases, when study laboratories chose a spike level that met all 4 downspiking criteria, either during the initial or the respiking phase, that level was sometimes well above two times the DL. Descriptive statistics of the QL_{DLT}/DL_T and QL_{DLK}/DL_K ratios for Laboratories 1 and 2 are shown in Table 5-6. (Statistics for Laboratory 3 are not shown because Laboratory 3 selected spike levels that were 2x the DL due the misunderstanding explained above.) In cases where the initial QL level was less than 2x the DL, the ratio was calculated using the respiked QL level; this is the ratio that is presented in Table 5-6. Additionally, analytes for which the initial spike levels did not meet MQOs were excluded from the calculation, as it can be concluded that the spike level was not too low in these cases.

Table 5-6. QL/DL Ratios for Laboratories 1 and 2

Laboratory	Limit	Geometric Mean Ratio of all Analytes	Minimum Ratio of all Analytes	Maximum Ratio of all Analytes
1	DL_T	3.77	2.00	9.64
	DL_K	4.24	2.23	9.44
2	DL_T	4.25	2.36	9.45
	DL_K	2.65	1.59	5.13

The chosen spike levels were frequently well above 2x the Task 1 detection limit for both DL_T and DL_K for Laboratory 1. This laboratory spiked at two different levels during the initial Task 2B phase. Although the second level was frequently the only one that exceeded 2x DL_K, the differences between the two spike levels was often much larger than would be necessary to meet the requirement for the two DLs. For example, while the two DLs for arsenic differed by a factor of less than two, the two QL spike levels differed by a factor of 5, and as a result, there was a much greater amount of distance between DL_K and QL_{DLK} than between DL_T and QL_{DLT}.

For Laboratory 2, the distance between DL_T and QL_{DLT} tended to be larger than the distance between DL_K and QL_{DLK}. This laboratory ran one spike level per analyte during the Task 2B initial phase. For some analytes, respiking was necessary to meet the 2x DL_K requirement but not the DL_T requirement. As a result, QL_{DLK} was greater than QL_{DLT} for these analytes; however the QL_{DLT}/DL_T and QL_{DLK}/DL_K ratios were similar. For other analytes, the initial spike level chosen met the 2x DL requirement for both DLs, and as such QL_{DLK} and QL_{DLT} were equal. For these analytes, the QL_{DLT}/DL_T ratio exceeded the QL_{DLK}/DL_K ratio.

The spiking scheme for Task 3 was designed to help address the issues noted above. For example, Laboratory 2's second QL_{DLK} spike level for barium still did not meet the 2x DL_K requirement, so EPA included a level that was approximately 2x DL_K for this analyte in Laboratory 2's Task 3 spiking scheme.. Likewise, EPA selected Task 3 spiking levels for Laboratories 1 and 2 that were below the QL for analytes for which the Task 2 QL/DL ratio was high. This allowed EPA to use Task 3 data to assess whether the laboratories' spike level choice

for those analytes was inappropriately high. This assessment is presented Section 7 of this report.

5.2.3.2 Downspiking Assessment - Method 625

The initial QL spike levels chosen by the three Method 625 laboratories frequently did not reflect the results of the downspiking analyses. For many analytes, the downspiking results seemed to indicate that the laboratories could have chosen a lower spike level than the ones that were ultimately used in the initial QL analyses.

When asked why these spike levels were chosen, the laboratories replied as follows:

- One laboratory stated that they believed they would be able to spike at a lower concentration than the initial-phase QL level during the respiking phase of Task 2C.
- One laboratory stated that initial-phase QL spiking was done at multiple levels for each analyte, however, the downspiking-determined level for all of the analytes was not included among these spike concentrations for all analytes.
- One laboratory stated that the goals of the downspiking could have been more clearly stated in the Statement of Work (SOW), but that downspiking levels below the initial QL failed the FACDQ procedure criteria in most cases.

The laboratory that assumed respiking could be done at lower levels in Task 2C misinterpreted the SOW. Respiking at lower levels is not required in the FACDQ 2.4T procedure and, therefore, was not included in the study. However, that laboratory also stated that qualitative identification was frequently not achievable at the downspiking level below the initial QL spike level, and therefore, the QLs for that laboratory likely would not have been different had downspiking been performed because the downspiking samples would not have met the FACDQ QL requirements for most analytes.

Because all laboratories also analyzed samples at seven or more spike levels as part of Task 2A, there were often multiple levels available that fell between the initial QL spike level and the optimal level that would have been chosen based on the downspiking analyses. Although Task 2A required only four replicates per level rather than the seven required for Tasks 2B and 2C, Task 2A results from levels falling below the initial QL level offer an indication of whether the FACDQ QL MQOs could have been achieved at a lower concentration for Method 625. Table 5-7 shows the number of spike levels between the downspiking level and the initial QL spike level.

Table 5-7. MQO Assessment at Task 2A Concentrations Below Task 2C QL

Laboratory	Number of Task 2A Spike Levels below Initial QL Spike*	Number of Analytes with Task 2A Spike Levels below Initial QL Spike	Number of Analytes Meeting MQOs at Lower Level
4	0	0	0
5	1-5	49	47
6	0-1	47	0

*A range indicates that the number of spike levels varies by analyte

Among the three laboratories, only Laboratory 5 appeared to have generated FACDQ QLs that did not reflect the lowest concentration that may have been capable of meeting the MQOs. Based on the Task 2A data evaluation, Laboratory 5's QLs could have been as much as 6 times lower than the value that was actually chosen by the laboratory. Although Laboratory 5's temporal range and batch frequency within each Task 2A spike level was comparable to that of their Task 2C data, the smaller number of replicates required in Task 2A could have an effect on the laboratory's ability to meet the MQOs at these lower levels. Therefore, the spike levels assigned to Laboratory 5 for Task 3 included at least one spike level below the QL for most analytes. This allowed EPA to assess whether the laboratory's originally determined QLs could have been lower.

For Laboratory 4, no Task 2A spike levels were run below the lowest of the three spike levels run initially for Task 2C, and for Laboratory 6, the RSD MQO could not be met for any analytes at the Task 2A level below the initial Task 2C level.

5.3 MRL Determination

Once LCMRLs were calculated for each analyte/laboratory set, EPA determined an MRL for each analyte. For consistency, only the spike levels used in the final LCMRL calculation were included in the MRL calculation (i.e., extra spike levels or replicates that were submitted by the laboratory were excluded). Following the MRL algorithm, an MRL could be determined for all 24 Method 200.7 analytes, and for 46 of the 49 Method 625 analytes. The three analytes for which an MRL could not be determined (2,4-Dimethylphenol, 2-Nitrophenol, and 3,3'-Dichlorobenzidine) also were the only analytes for which only one of the three laboratories could determine an LCMRL.

5.4 Assessment of Task 2 Temporal and Batch Variability

In Task 2 of the study, laboratories were instructed to include a specified amount of temporal variability among their analyses. This was done to simulate the routine variability that would be expected when running these procedures in practice, without conducting the study over an impractically long period of time. For example, the FACDQ procedure specifies that laboratories use data generated in routine practice over the course of a full year to calculate their FACDQ detection and quantitation limits, but conducting this study over a full year was not practical.

To address the need for temporal variability, laboratories were instructed to do the following when spiking and analyzing samples to be used to calculate LCMRLs:

> *"Prepare and analyze four replicate samples at each selected spike level. For each method, these replicates must be analyzed over at least three different preparation/analysis batches for each concentration (i.e., there must be at least 3 prep batches associated with each spike level). These three preparation/analysis batches must be spread out over approximately two weeks. Additionally, each laboratory performing Method 625 must perform a new initial calibration in the middle of this two-week analysis period. (Method 200.7 requires that a single-point calibration be performed daily.)"*

Labs were instructed to do the following when spiking and analyzing samples to be used to calculate and evaluate FACDQ QLs for Method 200.7:

> *"Prepare and analyze 7 replicate samples at the selected FACDQ QLDL_T and QL DL_K spike levels. These replicates must be analyzed over at least three different preparation/analysis batches for each concentration. In other words, labs may NOT prepare and analyze all 7 replicates for the QL_{DLT} in only 1-2 batches, nor may labs analyze all 7 replicates for the QL_{DLK} in only 1-2 other batches. The three preparation/analysis batches for each set of QL determinations must be spread out over approximately two weeks."*

Labs were instructed to do the following when spiking and analyzing samples to be used to calculate and evaluate FACDQ QLs for Method 625:

> *"Prepare and analyze 7 replicate samples at the FACDQ 2.4T spike level selected for each analyte. These replicates must be prepared and analyzed over at least three different preparation/analysis batches (i.e., analyzing all 7 replicates in only one or two batches is prohibited). These three preparation/analysis batches must be spread out over approximately two weeks. Additionally, the laboratory must perform a new initial calibration in the middle of this analysis period (corresponding to the same initial calibration described in SOW Section 3.2.1 Step 2 for determination of the OGWDW LCMRL."*

During the Task 2 data review process, each laboratory's data were assessed to make sure they met each of the following requirements:

- Results for each sample type (i.e., FACDQ QL_{DLT} spike, FACDQ QL_{DLT} spike, and LCMRL spike), analyte, and spike level were spread across at least three preparation batches.
- Results for each sample type, analyte, and spike level were spread across at least three analysis batches.
- The last replicate within a sample type, analyte, and spike level was prepared approximately two weeks after the first replicate within that sample type, analyte, and spike level.
- The last replicate within a sample type, analyte, and spike level was analyzed approximately two weeks after the first replicate within that sample type, analyte, and spike level.
- Results for each sample type, analyte, and spike level were associated with at least two calibrations.

For the purpose of the review, "approximately two weeks" was interpreted to mean that the minimum acceptable date range between the first and last replicate within a sample type, analyte and spike level was 11 days. (An 11 day period equates to beginning analysis on Monday during the first week and ending analysis on Friday of the second week.)

Results of this review indicated that the laboratories failed to meet all of the above criteria for some or all of the Task 2 data. Each laboratory was subsequently contacted to determine

whether this was due to misinterpretation of the instructions, practical issues such as scheduling, or other reasons. A discussion of each of the temporal requirements is provided in Sections 5.4.1 – 5.4.4 below.

During the discussion of the temporal requirements, Laboratory 5 suggested that they could combine some of their Task 2A and 2C data, which would create modified sets of replicates that meet the calibration and batch requirements, and increase the date ranges from what had originally been submitted. The evaluation of this laboratory's data focused on the modified data rather than the original submission, however responses to questions about their original interpretation of the SOW and procedure instructions also are discussed in the sections below.

5.4.1 Calibration Requirements

Most laboratories performed the required recalibration during their Task 2 analyses, with the following exceptions:

- Laboratory 1 did not include a recalibration within the lowest respiking level performed for Task 2A.
- When they initially submitted their data, Laboratory 5 did not include a recalibration within each spike level for their initial Task 2C analyses, or for their Task 2A or Task 2C respiking analyses at one level (50 ug/L). The resubmitted data included the recalibration within all initial Task 2C analyses, but not the respiking analyses.

EPA does not believe these deviations had a significant impact on the overall study results. Laboratory 1 spiked only three analytes (beryllium, cadmium, and barium) at the lowest Task 2A level. The laboratory had already respiked at two additional spike levels for these three analytes, and those additional spike levels did include the required recalibration. As a result, the laboratory's LCMRL calculation for these three analytes included 9 spike levels for which the recalibration was run, and 1 spike level for which it was not. Therefore, it is unlikely that the missing Laboratory 1 recalibration would have had a strong effect on the calculated LCMRLs for these three analytes. To mitigate the effects of any low bias that might have been caused by the missing recalibration, however, EPA selected Task 3 spike levels for Laboratory 1 that were greater than the calculated LCMRLs for all 3 analytes. (Results of EPA's assessment based on the Task 3 spike levels are given in Section 7 of this report.) As noted above, EPA also was able to overcome Laboratory 5's misunderstanding of the calibration requirements by combining data generated during Tasks 2A and 2C. Therefore, the missing recalibrations within each task did not have any affect on Laboratory 5's LCMRL or their FACDQ limits.

5.4.2 Batch Requirements

The number of preparation and analysis batches run by each laboratory for each task and round of spiking are presented in Table 5-8 below.

Table 5-8. Task 2 Laboratory Batch Frequencies

Method	Laboratory	Task	Spike Round	# of Spike Levels	# of Prep Batches	# of Analysis Batches	Were Requirements Met?
200.7	1	2A	Initial	7	2	3	No
200.7	1	2A	Respike (Sodium)	1	3	3	Yes
200.7	1	2A	Respike (Ba, Be, Cd)	3	2-3	1-3	No (for lowest level only)
200.7	1	2A	Respike (8 analytes)	2	3	3	Yes
200.7	1	2B	Initial	2	2-3	3	Yes for one spike level; no for one spike level
200.7	1	2B	Respike	1	3	2	No
200.7	2	2A	Initial	11	2-3	3-4	Yes for most spike levels, no for 2 spike levels
200.7	2	2B	Initial	1	5	5	Yes
200.7	2	2B	Respike	1	3	3	Yes
200.7	3	2A	Initial	9	3	3	Yes
200.7	3	2A	Respike	2	3	3	Yes
200.7	3	2B	Initial	4	3	3	Yes
625	4	2A	Initial	13	3	2-3	Yes for most spike levels, no for 3 spike levels
625	4	2A	Respike	1	3	3	Yes
625	4	2C	Initial	3	3	3	Yes
625	4	2C	Respike	2	3	3	Yes
625	5	2A	Initial	10	4	2	No
625	5	2A	Respike	2	3	1	No
625	5	2C	Initial	2	3-4	2*	No
625	5	2C	Respike (bis (2-Ethylhexyl) phthalate)	1	4	3	Yes
625	5	2C	Respike (3,3'-Dichlorobenzidine)	1	4	1	No
625	6	2A	Initial	7	4	3	Yes
625	6	2A	Respike	1	3	3	Yes
625	6	2C	Initial	1	3	3	Yes
625	6	2C	Respike	1	3	3	Yes

* Based on resubmission; Only one analysis batch in original submission.

Laboratories 3 and 6 met both the prep batch and the analysis requirements for all sets of analyses. Laboratories 2 and 4 met all batch requirements for all analyses with the exception of 2 to 3 of the Task 2A spike levels. Because of the large number of total spike levels run by these two laboratories during Task 2A (both laboratories ran more than the instructed 7 spike levels per analyte in this task), this departure was considered to be minor.

Laboratories 1 and 5 had more notable departures from the study's batch requirements. Laboratory 1 only included two preparation batches for the initial Task 2A samples and for one

of the two spike levels initially run for Task 2B. Laboratory 1's respiking for Task 2B only included 2 analysis batches, and respiking for Task 2A included as few as a single analysis batch for a given spike level and analyte. Laboratory 5 included the appropriate number of preparation batches, but only included 1analysis batch per spike level for the respike analyses.
When asked about why the batch requirements were not met, labs gave various explanations, including:

- Two laboratories stated that they interpreted 'analysis batch' to be synonymous with 'preparation batch' and, therefore, the sample preparation but not the sample analysis was spread over the required number of batches. EPA believes this is an unusual interpretation in that most analytical methods differentiate between preparation batch QC requirements (e.g., preparation blanks) and analysis batch QC requirements (e.g., calibration verification standards).
- Multiple laboratories stated that the schedule, as well as difficulties with the LCMRL calculator software and other practical laboratory issues, forced the batching and timeframe to be compressed in order to meet the required deadlines.

It is also worth noting that much of the data for which the batch requirements were not met were extra spike levels beyond what was required in the SOW. For example, Laboratory 4 ran 13 spike levels initially for Task 2A with 10 meeting the batch requirements, and Laboratory 1 respiked at 3 levels per analyte for Task 2A rather than 1, with only the third spike level including just a single analysis batch.

5.4.3 Temporal Requirements

The number of days across which sample preparation and analyses were spread by each laboratory for each task and round of spiking is presented in Table 5-9 below.

Table 5-9. Task 2 Laboratory Temporal Ranges

Method	Laboratory	Task	Spike Round	# of Spike Levels	Temporal Spread (Days)		Were Requirements Met?
					Preparation	Analysis	
200.7	1	2A	Initial	7	9	9	No
200.7	1	2A	Respike (Sodium)	1	12	11	Yes
200.7	1	2A	Respike (Ba, Be, Cd)	3	3-4	1-4	No
200.7	1	2A	Respike (8 analytes)	2	3	4	No
200.7	1	2B	Initial	2	3-5	5-7	No
200.7	1	2B	Respike	1	6	3	No
200.7	2	2A	Initial	11	2-3	2-3	No
200.7	2	2B	Initial	1	15	10	No
200.7	2	2B	Respike	1	13	8	No
200.7	3	2A	Initial	9	4	12-48	No
200.7	3	2A	Respike	2	3	5-7	No
200.7	3	2B	Initial	4	7	4-10	No
625	4	2A	Initial	13	3-6	2-4	No
625	4	2A	Respike	1	10	11	No
625	4	2C	Initial	3	11	10	No

Table 5-9. Task 2 Laboratory Temporal Ranges

Method	Laboratory	Task	Spike Round	# of Spike Levels	Temporal Spread (Days)		Were Requirements Met?
					Preparation	Analysis	
625	4	2C	Respike	2	7	7	No
625	5	2A	Initial	10	10	8	No
625	5	2A	Respike	2	3	1	No
625	5	2C	Initial	2	9-14*	8*	No
625	5	2C	Respike (bis (2-Ethylhexyl) phthalate)	1	36	34	Yes
625	5	2C	Respike (3,3'-Dichlorobenzidine)	1	8	1	No
625	6	2A	Initial	7	15	12-13	Yes
625	6	2A	Respike	1	10	5	No
625	6	2C	Initial	1	12	13	Yes
625	6	2C	Respike	1	6	4	No

* Based on resubmission; original submission included only a 5-day preparation range and a 1-day analysis range.

No laboratory met both the preparation and the analysis day requirements for all samples, and only three laboratories met all the preparation and analysis day requirements for any samples (Laboratory 6 met the requirements for all initial-phase samples for both tasks, Laboratory 1 met the requirements for Task 2A respiking of sodium only, and Laboratory 5 met the requirements for Task 2C respiking of bis (2-ethylhexyl) phthalate only).

In a few cases, the deviation from the required temporal range was only slight. Laboratory 2's initial-phase Task 2B samples and Laboratory 4's initial-phase Task 2C samples were prepared over the required amount of time, but missed the 11-day cutoff for sample analysis by only one day. Laboratory 4's Task 2A respiking phase samples made the 11-day cutoff for sample analysis, but missed sample preparation cutoff by one day. Laboratory 2's respike-phase Task 2B samples also met the sample preparation time requirement, but missed the analysis time requirement by 3 days. Laboratories 1 and 5 both spread their sample preparation and their sample analysis of initial-phase Task 2A samples over more than one week but less than the minimum 11 days.

In all other cases, sample preparation and/or sample analysis was spread over one week or less for at least some of the analytes. When asked about why the temporal requirements were not met, labs gave various explanations, including:

- Several laboratories expressed confusion regarding the SOW instructions for the required time frame. Multiple laboratories cited the phrase "approximately two weeks" and stated that they interpreted this to mean that:
  - time frames as short as 8 days would meet this requirement,
  - the phrase merely meant that preparation and analysis must not be done on consecutive days, or

  - the phrase referred to calendar week (i.e., the sampling and analysis must be spread across multiple calendar weeks, but the actual number of days between beginning and completion did not matter).
- One laboratory stated that they assumed that the two-week time period included the downspiking analyses as well as the initial Task 2B/2C analyses
- One laboratory stated that they understood the two-week requirement to be a maximum rather than a minimum requirement.
- Multiple laboratories stated that the schedule, as well as difficulties with the LCMRL calculator software and other practical laboratory issues, forced the batching and timeframe to be compressed in order to meet the required deadlines.

5.4.4 Effect of Compressed Variability

Although data that did not meet the temporal and batch requirements of the study could easily be identified, it was not as clear what affect the compressed timeframe or batching would have on the variability of those data, or how they would affect the limits determined using the data. There were a few cases in which the same analyte/spike level combination was included in both the LCMRL and FACDQ procedure evaluations. In all of these combinations, at least some of the temporal/batch requirements were missed for both the LCMRL and FACDQ data; however one of the two sets missed the requirements by a greater amount than the other. When comparing these data, the set with greater temporal spread did not consistently yield greater variability than the set with lower temporal spread; however, EPA decided that the laboratories should be instructed to meet the temporal requirements (with the instructions more clearly described) when performing sample preparation and analysis during Task 3. This was done to avoid bias in the comparisons between the Task 2 limits and the Task 3 data.

Additionally, when responding to the temporal/batch questions, Laboratory 5 indicated that their initial Task 2C spike levels overlapped with two of the Task 2A spike levels, and therefore, these data could be combined to increase the temporal period associated with each Task 2C spike level. The laboratory provided revised calculations, with the last two Task 2C replicates analyzed for each spike level replaced with Task 2A and Task 2C downspiking results that had been prepared prior to the Task 2C replicates. By doing this, the re-calibration requirement would be met, the preparation time range requirement would be met for one spike level and would be increased to nine days for the other, and the analysis time range and number of batches would increase to 10 days and 2 batches, respectively. This change did not affect whether the MQOs were met at the spike levels, though they did have slight effects on the resulting DLs. The DLs and QLs determined based on the revised calculations submitted by this laboratory were the limits used in the MQO assessments presented in Section 6 and 7.

Section 6: DL Assessment – Method 625

As described in Section 3.4, detection limits were calculated for each Method 625 analyte/lab set following the steps and formula in the FACDQ 2.4T procedure for censored methods. A discussion of how various steps in the FACDQ 2.4T procedure affected the resulting limits is presented in Section 6.1.

As with the FACDQ DL for uncensored methods, the censored method DL was described as targeting a ≤1% false positive rate in the FACDQ procedure. Therefore, EPA compared the calculated FACDQ DLs to the target 1% rate, using the blank data submitted by the laboratories in Task 1 and the preparation blanks run with each batch during Tasks 2 and 3. This comparison is presented in Section 6.2. Assessments of how various assumptions made in the procedure, including the FACDQ procedure method for classifying methods as censored or uncensored based on blank data, distribution assumptions regarding the blank results, and the effect of outlier testing on the resulting limits, affect the resulting rate are presented in Section 6.3.

6.1 DL Calculations

Unlike the DLs calculated for Method 200.7, the Method 625 FACDQ DLs were calculated using spiked sample data. Specifically, the initial FACDQ QL spike level that met the downspiking requirements was used to calculate the DL. Per the FACDQ procedure requirements, it was not necessary for this spike level to meet the QL MQOs, and the DL was not recalculated using higher spike level data when respiking was required to meet those MQOs. Once the FACDQ DL was calculated, the limits were compared to the Task 1 blank data and adjusted, where necessary, based on the FACDQ 2.4T ongoing verification methodology. To accomplish this, the highest non-outlying blank result among the first 20 blanks analyzed was identified for each analyte/laboratory set. This ongoing verification assessment was performed in the same way as described in Section 4.1, with a maximum of one blank result removed for each analyte/lab set.

At least one outlier was identified and removed in 48 of the 147 analyte/lab sets of 20 blanks. Outliers were identified most frequently for Laboratory 4, which tended to have the highest rate of detects in the Task 1 blank results. The majority of outlying blank results were either PAHs (27 outlying results) or phthalates (15 outlying results). This is not surprising, as these analytes also were most frequently detected in blanks during Tasks 2 and 3 for Method 625.
Once the outliers were removed, the highest remaining result among the 20 blanks was identified for each analyte/lab set. This blank result exceeded 0 in 44 of the sets, most frequently for Laboratory 4, and most frequently for PAHs and phthalates.

The FACDQ DL calculated using the spiked sample results rarely needed to be adjusted based on the blank results. The highest non-outlying blank result exceeded the DL calculated from the spiked sample results for three analyte/lab sets [benzo(b)fluoranthene and benzo(g,h,i)perylene for Lab 4; di-n-butyl phthalate for Lab 5]. The DLs for these three sets were increased by 20.6 – 96.3%. Overall, the FACDQ DL determined from spiked sample results exceeded the highest non-outlying DL by a median of 6.4 times and a geometric mean of 8 times, among the 44 sets for which the highest blank was a detect.

6.2 MQO Assessment

Each laboratory submitted between 30 and 32 blank results per analyte during Task 1. After excluding the 20 blanks used in the DL adjustment, only 10-12 blanks per analyte/lab set remained. To increase statistical power when assessing the Method 625 DLs, EPA supplemented the remaining Task 1 blanks with preparation blanks that were analyzed as routine QC during Tasks 2 and 3. Because this assessment was to be made using all analytes and laboratories combined, EPA excluded blanks that were analyzed during the respiking phase of Task 2 to avoid biasing the assessment with analytes that had trouble meeting the LCMRL and FACDQ MQOs. This resulted in a set of data with between 19 and 27 blank results per analyte and laboratory. As with the Method 200.7 blanks, this dataset was called the "full verification set."

EPA compared the full verification set of blanks to the calculated DL values for each analyte/ laboratory set, with any blank result exceeding the DL being categorized as a false positive result for that limit. EPA then calculated the percentage of false positives over all analytes and laboratories. EPA also calculated the false positive rate after applying the FACDQ 2.4T procedure's outlier test to the full verification data set. The overall false positive percentages for the DLs with and without outlier removal are presented in Table 6-1.

Table 6-1. Method 625 False Positive Rates for the FACDQ DL, Full Verification Dataset

Outliers Removed in Full Verification Data?	Number of Total Blank Results	Number of False Positives	Percent of False Positives
No	3,379	18	0.53
Yes	3,324	2	0.06

The 18 false positives observed in the full verification datasets prior to outlier removal included 10 PAH results and 8 phthalate results. Only bis (2-ethylhexyl) phthalate, butyl benzyl phthalate, and di-n-butyl phthalate had more than one false positive. After outlier removal, only two false positives remained, including one blank result each for butyl benzyl phthalate and di-n-butyl phthalate.

Although an overall false positive rate can be calculated using all blank results, it is possible that this estimate could be influenced by correlations between analytes. Because factors that would influence whether a blank result is low or high would likely affect more than one analyte, combining all blank results from all analytes to determine a single rate and compare it statistically to the target 1% rate would potentially yield biased results. Therefore, EPA determined a bootstrap estimate of the false positive rate by randomly selecting 100 sets of blank results for each analyte/lab set. A similar data analysis was performed on Method 200.7 results, as described in Section 4.2. Because there were a smaller number of blanks per analyte, but a larger number of analytes for Method 625 compared to Method 200.7, each of the 100 sets included 10 blanks selected with replacement (i.e., the same blank result could be selected more than once), rather than 20 blanks, per analyte and lab. As with the Method 200.7 assessment, the bootstrap estimate of the false positive rate for Method 625 was the mean of the 100 false positive rates calculated for that analyte/lab. EPA compared this estimate to the target 1% using a one-sample proportion test following the Binomial distribution, run at the 95% confidence level. The bootstrap-estimated false positive rate for the DL and the result of the proportion test are presented in Table 6-2.

Table 6-2. Method 625 False Positive Rates, Bootstrap Estimates

Outliers Removed in Full Verification Data?	Number of Total Blanks per Bootstrap Run	Mean False Positive Rate	p-Value for Proportion test
No	1,470	0.49%	0.021
Yes	1,470	0.07%	<0.0001

A likely reason for the low false positive rates observed for Method 625 is that many analytes never were detected in blanks. Because no non-zero DL could accurately estimate the target MQO for these analytes, EPA also calculated the false positive rate among only those analyte/lab sets for which a measured result was observed in at least one blank sample. The results of this assessment are presented in Table 6-3.

Table 6-3. False Positive Rate Assessment – Sets with Blank Hits only

Outliers Removed?	Total Number of Blanks	False Positive Rate Based on Full Verification Set	Number of Total Blanks per Bootstrap Run	Mean False Positive Rate	p-Value for Proportion test
No	1,454	1.24%	600	1.20%	0.255
Yes	1,399	0.14%	510	0.19%	0.037

When only the analyte/lab sets with at least one blank hit were assessed, the overall false positive rate was 1.24%, and the bootstrap estimate for those analyte/lab sets was 1.20%, which was not significantly different from the target 1% at the 95% confidence level. However, the overall false positive rate and bootstrap-estimated false positive rates calculated after outlier removal were still lower than 1%, with rates of 0.14 and 0.19%, respectively.

6.3 Factors Affecting MQO Assessment

6.3.1 Censored/Uncensored Classification

For the purpose of the study, the FACDQ DLs for Method 625 were calculated based on the censored method/analyte calculations in the FACDQ 2.4T procedure. However, the procedure states that this designation should be made on an analyte basis using the results of blank analyses. Therefore, had the assumption that all analytes met the "censored" classification not been made, it is possible that some of the FACDQ DLs would have been calculated differently. To evaluate this assumption, EPA examined the frequency of blank samples that had hits among several sets of data, including all Task 1 through 3 blank samples, the 20 Task 1C blank samples that were used to assess the FACDQ DLs calculated in this study (as described in Section 6.1), and a set of 7 blank results per analyte/lab set randomly selected from the 3 Tasks (excluding the Task 1C blanks). The seven results were chosen randomly across the three tasks. This was done because differences in blank concentration were observed for several analytes between tasks, and therefore, unexpectedly high or low false positive rates could be due to systematic differences between the blanks used to calculate the DLs compared to the blanks used to assess the DLs, rather than to the calculations themselves producing inaccurate DLs.

EPA used the blank data described above to determine the frequency of analyte/lab sets that yielded at least 50% hits, as well as the frequency of sets that yielded hits exceeding a second cut-off (75%), which was suggested as part of an EPA-sponsored Peer Review of the FACDQ

2.4T procedure. The frequency at which analyte/lab sets were classified as uncensored based on the different datasets and criteria are presented in Table 6-4.

Table 6-4. Frequency of Method 625 Analytes Exceeding the Uncensored Method Classification

Blank Dataset	Percentage of Analyte/Lab Sets with at least 50% hits	Percentage of Analyte/Lab Sets with at least 75% hits
All Blanks	9.5	6.1
Task 1C Blanks	10.2	6.8
Random Selection of 7 Blanks	11.6	6.1

The frequency of analyte/lab sets that yielded at least 50% hits varied slightly between the different datasets. In all cases, the analytes identified were either phthalates or PAHs (with the majority being phthalates). The frequency of analyte/lab sets that yielded at least 75% hits varied somewhat less by dataset than the frequency for the 50% cut-off. In all cases, the analytes with at least 75% hits were phthalates.

To assess the effect of the censored/uncensored determination on the resulting DLs and false positive rates, EPA calculated four additional FACDQ DLs for lab/analytes sets yielding at least 50% hits and lab/analytes sets yielding at least 75% hits, as outlined below.

- DL_T according to the formula in the FACDQ 2.4T procedure and used in the 200.7 DL assessment
- DL_K according to the alternate formula used in the 200.7 DL assessment
- An adjusted DL_T, calculated using only the blanks that generated hits in the calculation
- An adjusted DL_K, calculated using only the blanks that generated hits in the calculation

These DLs were calculated using 7 randomly selected blanks, and then adjusted using the 20 Task 1C blanks and compared with Task 2 DLs. The latter two DLs listed above were calculated using only those blanks that generated hits; this calculation was suggested during the Peer Review of the FACDQ 2.4T procedure.

The adjusted DLs differ from the DLs calculated following the study methodology because the blanks that did not generate signals are excluded from the calculation instead of setting their results to 0. As a result, the adjusted DLs are calculated using a higher mean, a lower standard deviation, and higher multipliers than the unadjusted DLs.

EPA compared the four additional DLs, to the remaining blank results (i.e., excluding the 7 replicates used in the calculation and the 20 Task 1C blanks) to determine false positive rates. The FACDQ DLs that were calculated during Task 2 (as described in Section 6.1) also were included in this assessment. Overall false positive rates for those sets that met the 50% cutoff, and those sets that met the 75% cutoff, are presented in Table 6-5. The individual false positive rates for these sets are shown in Figures 6-1 and 6-2.

Table 6-5. False Positive Rates for Method DLs Calculated from Blanks

Limit	FP Rate (all sets meeting 50% blank hit cutoff)	FP Rate (all sets meeting 75% blank hit cutoff)
DL_T	1.66%	1.27%
DL_K	0.67%	0.64%
Adjusted DL_T	1.16%	1.27%
Adjusted DL_K	0.67%	0.64%
Task 2 DL	1.00%	1.91%

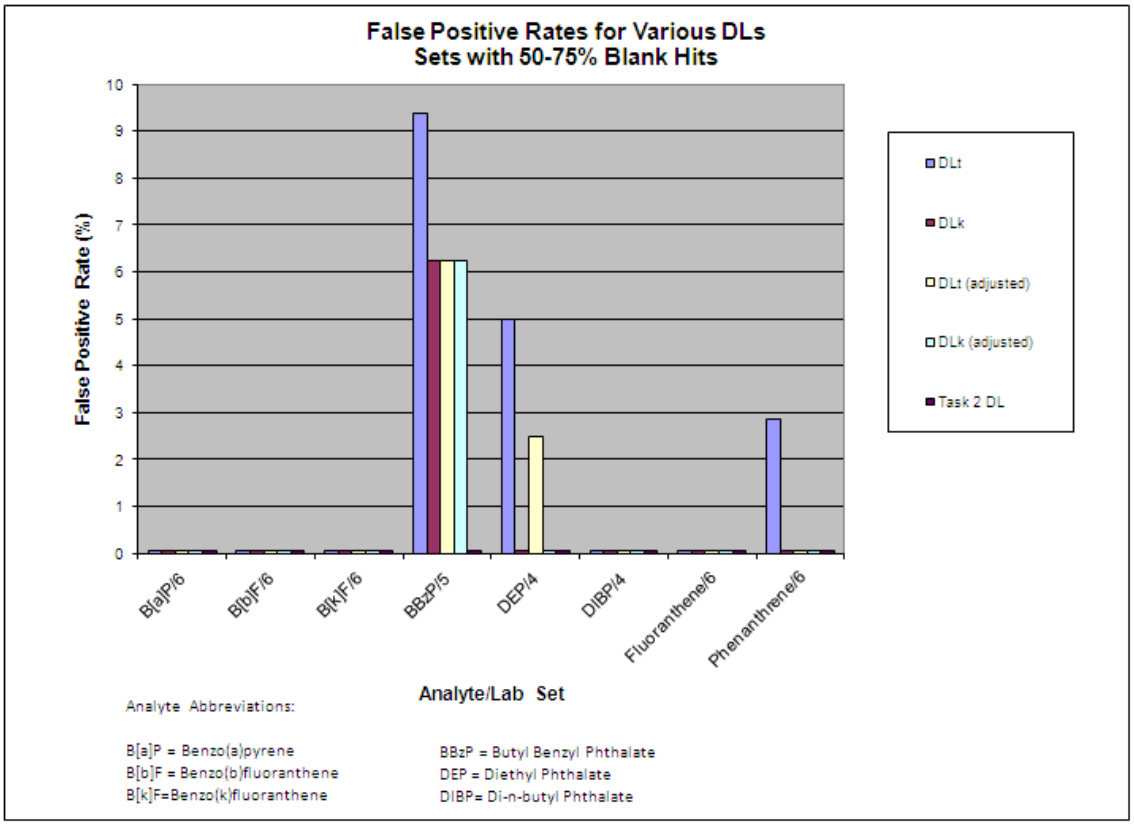

Figure 6-1. False Positive Rates for Various DLs, Sets with 50-75% Blank Hits

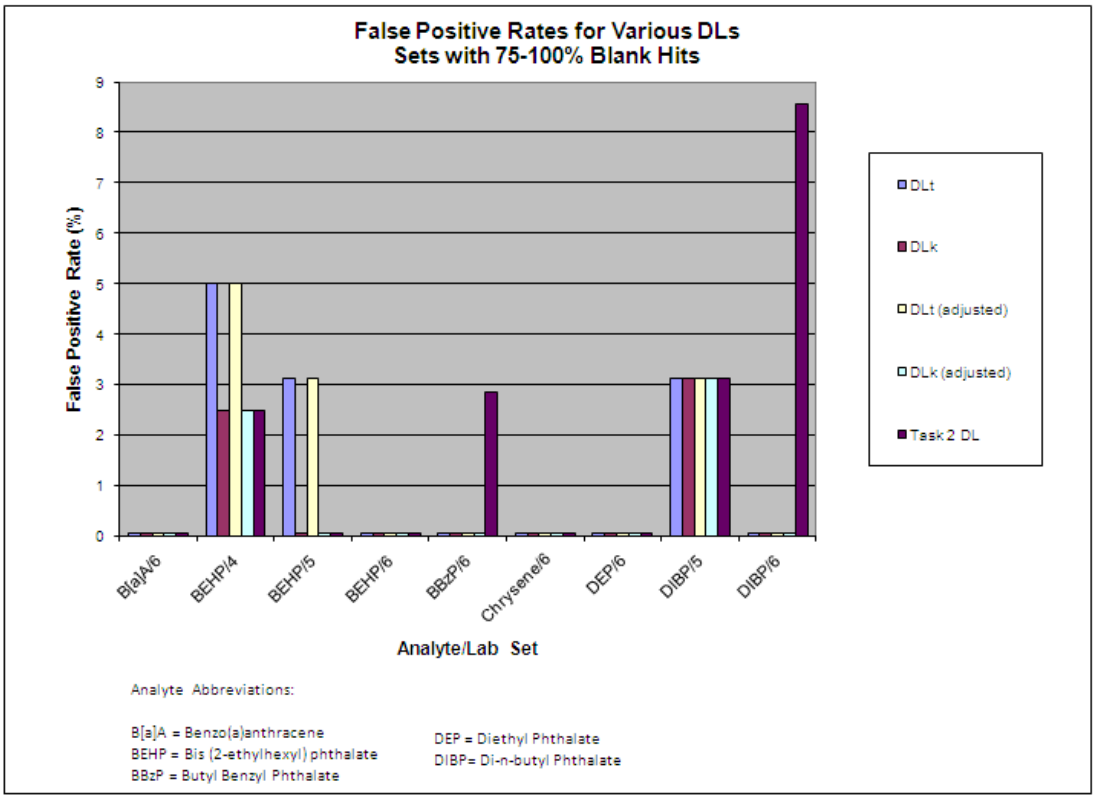

Figure 6-2. False Positive Rates for Various DLs, Sets with 75-100% Blank Hits

6.3.2 Distribution of Blanks

The distribution of blanks will be heavily affected by the frequency that an instrument signal was observed. If a large percentage of blank results did not yield an instrument signal, and a value of 0 was used as the numeric result for those blanks, the overall distribution will not be normal. EPA assessed the distribution of blank results for the 14 analyte/lab sets that generated hits across all blank results using the same methodology described previously (see Section 4.3). For the purpose of the distributional assessments, all previously identified outliers were included. For all 14 sets, the hypothesis that the blanks follow a normal distribution was rejected. In all cases, the blank distribution had a strong positive skewness. Figures 6-3 and 6-4 show the blank distribution for two of these sets [di-n-butyl phthalate for Lab 5, and bis (2-ethylhexyl) phthalate for Lab 4].

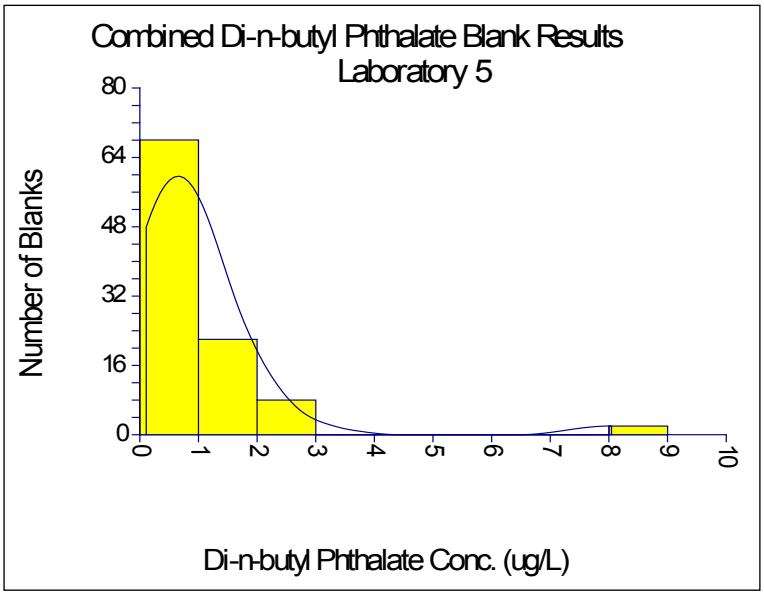

Figure 6-3. Distribution of Di-n-butyl Phthalate Blanks, Laboratory 5

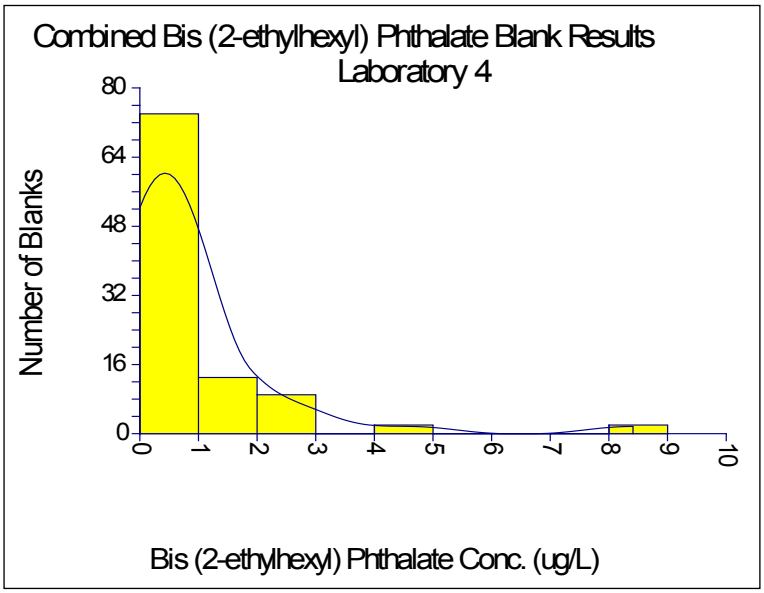

Figure 6-4. Distribution of bis (2-Ethylhexyl) Phthalate Blanks, Laboratory 4

6.3.3 Outlier Testing

As stated in Section 4.3.2, peer reviewers of the FACDQ 2.4T procedure suggested various alternative outlier tests to the one included in the procedure. EPA used the Task 1 blank data to assess the effect of these alternative tests on the identification of outliers and their effect on the FACDQ DL calculations. The alternative outlier tests included:

- Setting the upper and lower bounds to 3 standard deviations outside the mean.

- Setting the upper and lower bounds to the 97.5th percentile t-statistic outside the mean, with the degrees of freedom set to the number of blanks used in the assessment, minus 1.
- Setting the upper and lower bounds to the median ± 4*MAD, where MAD corresponds to the median of the absolute differences between the individual results and the median. This approach is less sensitive to departures from a normal distribution.
- Grubbs test, a frequently used outlier test

Each of these outlier tests was performed using just the Task 1C blank data, adhering to the FACDQ 2.4T rule that no more than 5% (i.e., 1 out of the 20 Task 1C blanks) may be removed per analyte/laboratory set. The frequency of Task 1C outlier removal based on the different tests is presented in Table 6-6 below.

Table 6-6. Frequency of Task 1C Outlier Removal, Method 625

Test	Number of Low Outliers Removed	Number of High Outliers Removed	Number of Total Outliers Removed
Mean ± 2SD (FACDQ 2.4T outlier procedure)	1	47	48
Mean ± 3SD	0	33	33
Mean ± tSD	1	46	47
Median ± 4MAD	1	48	49
Grubbs test	0	39	39

Unlike Method 200.7, almost all outliers identified for Method 625 were high outliers. The only low outlier identified was a single di-n-butyl phthalate result for Laboratory 4. The low outlier identified for this set was one of two blanks that did not generate a signal; the other 18 Task 1C blanks for this analyte yielded positive results.

As with Method 200.7, there was an inverse relationship between the number of outliers removed and the width of the bounds determined for the specific outlier test. For a set of 20 blanks, the multipliers for the mean ± t SD test and Grubbs test are 2.1 and 2.71, respectively. As a result, the number of analyte/laboratory sets with an outlier removed for the mean ± t SD test was close to that of the 2 SD test, and the number of analyte/laboratory sets with an outlier removed for Grubbs tests was close to that of the 3 SD test.

Unlike Method 200.7, the test identifying the largest number of outliers for Method 625 was the median ± 4MAD test. The MAD calculated based on the 20 Task 1C blanks was 0 for 89.8% of the analyte/lab sets. This occurred for any set that had hits in fewer than 25% of the blanks. Any hits observed for those sets would be identified as outliers, regardless of the numeric result. Of the 49 outliers identified based on the median ± 4MAD test, 35 occurred for sets for which the MAD was 0.

Though the number of outliers identified varied between the different tests, the ultimate impact was very slight, as shown in Table 6-7. Regardless of which outlier test was used, only 3 of the 147 DLs calculated during Task 2 fell below the highest non-outlying Task 1C blank and therefore would need to be adjusted. If no outliers were removed, 7 of the 147 DLs would be

adjusted. Therefore, not removing outliers had a net effect of yielding four slightly higher DLs and reducing the false positive rate based on the full verification set from 0.53% to 0.50%.

Table 6-7. Frequency of Task 1C DL Adjustment After Application of Various Outlier Tests

Outlier Test	% Analyte/Lab Sets With DL adjustment
None	4.8
Mean ± 2SD (FACDQ 2.4T outlier procedure)	2.0
Mean ± 3SD	2.0
Mean ± tSD	2.0
Median ± 4MAD	2.0
Grubbs test	2.0

6.3.4 Alternative Calculations

As stated in Section 4.3.4, peer reviewers commented that while the DL calculation in the FACDQ 2.4T procedure is considered to approximate a 99% upper prediction limit for a single sample analysis, it does not match the exact formula for a prediction limit. For censored methods/analytes, the exact calculation for a DL representing an upper 99% prediction limit would be:

$$DL_{PI} = t_{0.99;n-1)} * \sqrt{(1+\frac{1}{n})} * s$$

Where s is the standard deviation of the spiked sample results, and
n is the number of spiked sample results used in the calculation.

When calculated using 7 blank results, DL_{PI} would be approximately 7% greater than the DL calculated according to the FACDQ 2.4T formula. As with Method 200.7, EPA used the Method 625 blanks to determine the false positive rate for DL_{PI} and compared it to the false positive rate determined based on the FACDQ 2.4T DLs.

The calculated Method 625 DL_{PI} was 7% higher than the FACDQ 2.4T DL for 144 of the 147 analyte/lab sets. For the other three sets, the DL_{PI} and FACDQ 2.4T DL were equal; for these three sets, both the FACDQ 2.4T DL and DL_{PI} were adjusted based on the Task 1C blanks. EPA assessed the false positive rate for DL_{PI} using the full verification set of blanks described in Section 6.2, and compared the false positive rate to the target 1% using the same bootstrap estimation approach described in Section 6.2. The results are presented in Table 6-8.

Table 6-8. False Positive Rates for DL calculated using Exact Prediction Interval Formula

Set	Total Number of Blanks	False Positive Rate Based on Full Verification Set	Number of Total Blanks per Bootstrap Run	Mean False Positive Rate	p-Value for Proportion test
All Analytes/Lab Sets	3,379	0.50%	1,470	0.46%	0.021
Sets with at least one blank hit	1,454	1.17%	600	1.12%	0.255

Using the exact prediction limit formula had very little effect on the false positive rate. The overall false positive rate calculated using all analyte/lab sets dropped from 0.54% to 0.50%, and

the false positive rate calculated using only sets that generated at least one hit dropped from 1.24% to 1.17%. Similarly to the FACDQ 2.4T DLs, the bootstrap-estimated false positive rate was significantly different from 1% when all sets were included in the calculation, and were not significantly different from 1% when only sets with at least one blank hit were included.

Section 7: Task 3 LCMRL/FACDQ QL Assessments

EPA used a subset of the spiked replicate analyses performed during Task 3 to assess the FACDQ QLs and LCMRLs determined during Task 2. During Task 3, each laboratory was instructed to spike seven replicates at five different concentrations for each analyte. These five concentrations were chosen and assigned by EPA. Because some Task 2 limits tended to either vary by analyte or be values that were impractical for spiking (i.e., the calculated LCMRLs or FACDQ QLs were not round numbers), spike levels differed from the calculated Task 2 limits in some cases.

When selecting spike levels for each laboratory, EPA ensured that one spike level was at or close to the:

- LCMRL calculated during Task 2 (referred to as the "LCMRL assessment level" in this report)
- FACDQ QL determined in Task 2 for Method 625 (referred to as the "QL assessment level" in this report)
- Task 2 FACDQ QL_{DLT} for Method 200.7 (referred to as the "QL_{DLT} assessment level" in this report)
- Task 2 FACDQ QL_{DLK} when this limit differed from the QL_{DLT} (referred to as the "QL_{DLK} assessment level").

In some cases, EPA selected levels that were either slightly above or slightly below the limits determined in Task 2. This was typically done to address spiking practicalities in the laboratory or when the QL was adjusted based on the LER determined in Task 2. For Laboratory 3, the Method 200.7 QLs determined during Task 2 were exactly 2 times the corresponding DLs, and therefore were not round numbers for most analytes. Although this laboratory was able to spike at these levels during Task 2, it was impractical for them to re-spike at those levels while also spiking at the other levels needed to assess limits. Therefore, all the QL_{DLT} and QL_{DLK} assessment levels for Laboratory 3 were slightly above and slightly below their corresponding Task 2 QL limits. When selecting LCMRL, QL, $QL_{DLT,}$ and QL_{DLK} assessment levels that were slightly above or slightly below the corresponding limits determined in Task 2, EPA ensured that an approximately equal number of the Task 3 assessment levels were above and below the Task 2 limits.

Once the Task 3 analyses were completed by the laboratories and the data were submitted to EPA, EPA compared the Task 2 limits to the corresponding limit assessment data to assess whether the limits accurately estimated the minimum concentration to meet the procedure and method-specific MQOs. For analyte/lab sets for which an LCMRL or FACDQ QL could not be determined, no data from that set were included in the procedure-specific MQO assessments. For Laboratories 5 and 6, one replicate each at one spike level was determined to be invalid due to spiking issues identified by the laboratory. These replicate results were excluded from all analyses described in this section. All other replicates were considered valid, and were included in all analyses for which the spike level was considered applicable. Results of the LCMRL assessments are presented in Section 7.1, and results of the FACDQ QL assessments are presented in Section 7.2.

7.1 LCMRL Assessment

7.1.1 Comparison to Target MQO

For each analyte/lab set, EPA calculated percent recoveries for each of the valid replicate results at the LCMRL assessment level. The percentage of recoveries falling outside 50-150% was counted across labs and analytes for each method. Based on the procedure's target MQO, this percentage is expected to be 1%. Therefore, the observed frequency of recoveries outside 50-150% was compared to the target frequency based on a Binomial test, run at the 95% confidence level. The observed frequencies and test results for each Method are presented in Table 7-1.

Table 7-1. Overall LCMRL MQO Failure Rates

Method	Number of Calculated LCMRLs	Number of Evaluated Replicates	% Replicates outside 50-150%	p-Value of Binomial test
200.7	70	490	7.1%	<0.0001
625	119	802	2.9%	<0.0001

The observed frequency of replicates with recoveries outside 50-150% significantly exceeded 1% for both analytical methods. For Method 200.7, the majority of these replicates had recoveries exceeding 150%, while for Method 625, the majority of these replicates had recoveries below 50%. For both methods, p-Values below 0.05 indicate that the observed MQO failure rate differed significantly from 1% at the 95% confidence level.

Table 7-2 shows the frequency of LCMRL assessment results for each laboratory. When assessed separately by laboratory, the Task 2 LCMRLs yielded failure rates that did not differ significantly from the target 1% for two laboratories (2 and 5), that were significantly lower than 1% for one laboratory (4), and were significantly greater than the target 1% for three laboratories (1, 3, and 6).

Table 7-2. Lab-Specific LCMRL MQO Failure Rates

Method	Laboratory	Number of Calculated LCMRLs	Number of Evaluated Replicates	% Replicates outside 50-150%	p-Value of Binomial test
200.7	1	23	161	16.2%	<0.0001
	2	24	168	0.6%	0.314
	3	23	168	5.0%	0.0002
625	4	25	175	0%	<0.0001
	5	45	286	1.1%	0.224
	6	49	341	5.9%	<0.0001

7.1.2 Assessment of LCMRL MQO Deviations

As stated above, the accuracy of the calculated LCMRLs in estimating the minimum concentration to achieve the procedure's MQO appeared to vary widely by laboratories. This result is somewhat surprising, because all laboratories used the same automated LCMRL software, and the same amount of data was used for all laboratories (4 replicates each at 7 spike levels, with additional data added only if an LCMRL could not be determined initially).

However, as discussed in Section 5, not all laboratories spread out their analyses over the required numbers of batches and days during Task 2A. After Task 2 was completed, the temporal requirement instructions were revised for improved clarity, and all laboratories spread their sample preparation and analyses over the required number of batches and days. If the increased temporal spread increased the temporal variability in Task 3 relative to Task 2, it would be expected that the LCMRL MQO failure rate would exceed the target 1%. To examine this possibility, EPA compared the Task 2 temporal spread for each laboratory to the LCMRL MQO failure rates calculated from the Task 3 data. This comparison is presented in Table 7-3.

Table 7-3. Task 2 Temporal Spread Compared to Task 3 LCMRL MQO Failure Rates

Method	Laboratory	Task 2A Temporal Information (Initial Spike Levels only)				LCMRL MQO Failure Rate
		# Preparation Batches	# Analysis Batches	Prep Day Range	Analysis Day Range	
200.7	1	2	3	9	9	16.2%
	2	2-3	3-4	2-3	2-3	0.6%
	3	3	3	4	12-48	5.0%
625	4	3	2-3	3-6	2-4	0%
	5	4	2	10	8	1.1%
	6	4	3	15	12-13	5.9%

Because all laboratories spread Task 3 preparation and analyses across the required numbers of batches and days, the laboratories with the smallest temporal spread during Task 2 would be expected to have the greatest difference in variability between the two tasks. Due to this increase in variability, it would be expected that these laboratories would exhibit the highest LCMRL MQO failure rates. However, as can be seen in Table 7-3, this was not the case. Laboratories 1 and 6 had the highest LCMRL MQO failure rates, but also spread Task 2A preparation and analysis across a greater numbers of days than the laboratories with lower LCMRL MQO failure rates. Additionally, Laboratory 4 spread Task 2 sample preparation and analysis over a smaller number of days than most other laboratories, but produced a 0% LCMRL MQO failure rate in Task 3. One factor that could have influenced the LCMRL MQO failure rate was the number of preparation and/or analysis batches included in Task 2A. Even though the amount of time that their Task 2A and 3 analyses were spread across did not differ greatly, Laboratory 1 was the only laboratory that did not include 3 separate preparation batches in Task 2 for any analyte. It is possible that this affected the LCMRL MQO failure rate observed for this laboratory. However, it should be noted that Lab 6, which also had a high LCMRL MQO failure rate, included 4 separate preparation batches in Task 2, thereby casting doubt on this possibility.

Although deviations from the temporal requirements during the Task 2A initial phase preparation and analysis did not seem to affect the LCMRL MQO failure rates, differences in temporal spread between the initial phase and respiking phases of Task 2A may have. Table 7-4 lists all analyte/lab sets with at least one replicate falling outside 50-150% at the LCMRL assessment level, along with the amount of respiking necessary (if any) to determine a limit during Task 2A and the amount of temporal spread include in that respiking. Among the 30 analyte/lab sets for which at least one replicate did not meet the LCMRL MQO, 70% required respiking at one or more concentrations during Task 2, and among the 14 sets for which at least two replicates did not meet the LCMRL MQO, 79% required respiking during Task 2. For Laboratories 1 and 6, the Task 2A respiking concentrations were generally below the initial spike range, and the Task

2A respiking sample preparation and analysis was spread across fewer days than the initial spiking sample preparation and analysis. Therefore the level of temporal variability would have been the lowest at the lowest concentration(s); it is possible that this could have contributed to the high LCMRL MQO failure rates observed for these laboratories.

Table 7-4. Analyte/Lab Sets with LCMRL MQO failures compared to Task 2A Respiking Information

Method	Lab	Analyte	Number of Task 3 Replicates Outside 50-150%	Task 2A Respiking	# Batches in Respike Phase *	Day Range in Respiking Phase *
200.7	1	Aluminum	3	None	NA	NA
200.7	1	Barium	4	Below (3 levels)	2-3/1-3	3-4/1-4
200.7	1	Cadmium	1	Below (3 levels)	2-3/1-3	3-4/1-4
200.7	1	Chromium	5	Below (2 levels)	3/3	3/4
200.7	1	Cobalt	1	Below (2 levels)	3/3	3/4
200.7	1	Copper	3	Below (1 level)	3/3	3/4
200.7	1	Lead	1	Below (1 level)	3/3	3/4
200.7	1	Magnesium	1	None	NA	NA
200.7	1	Manganese	1	Below (2 levels)	3/3	3/4
200.7	1	Molybdenum	3	Below (2 levels)	3/3	3/4
200.7	1	Nickel	2	Below (2 levels)	3/3	3/4
200.7	1	Silver	1	Below (2 levels)	3/3	3/4
200.7	2	Manganese	1	None	NA	NA
200.7	3	Arsenic	1	None	NA	NA
200.7	3	Beryllium	2	None	NA	NA
200.7	3	Manganese	2	Below (1 level)	3/3	3/5
200.7	3	Tin	2	None	NA	NA
200.7	3	Zinc	1	None	NA	NA
625	5	2-Chlorophenol	1	Above (3 levels)	4/3	10/8
625	5	Diethyl phthalate	1	None	NA	NA
625	5	Nitrobenzene	1	None	NA	NA
625	6	2,4,6-Trichlorophenol	2	Below (1 level)	4/4	10/5
625	6	2,4-Dinitrophenol	1	None	NA	NA
625	6	2-Nitrophenol	2	Below (1 level)	4/4	10/5
625	6	4-Chloro-3-methylphenol	1	Below (1 level)	4/4	10/5
625	6	Benzo(ghi)perylene	1	Below (1 level)	4/4	10/5
625	6	bis(2-Ethylhexyl)phthalate	6	Below (1 level)	4/4	10/5
625	6	Butyl benzyl phthalate	1	Below (1 level)	4/4	10/5
625	6	Dibenzo(a,h)anthracene	2	Below (1 level)	4/4	10/5
625	6	*N*-Nitroso-di-*n*-propylamine	1	Below (1 level)	4/4	10/5
625	6	Pentachlorophenol	3	Below (1 level)	4/4	10/5

* values are for sample preparation/analysis

Respiking information was not presented in Table 7-4 for Laboratory 4 because no MQO failures were observed for that laboratory. Unlike Laboratories 1 and 6, Laboratory 4 performed respiking at a higher concentration than the initial range, with sample preparation and analysis

spread over a greater number of days than during the initial Task 2A spiking. This could have resulted in high-biased LCMRLs, which would be consistent with the 0% LCMRL MQO failure rate observed for this laboratory.

In addition to the possibility that higher LCMRL MQO failure rates could have been caused by differences in temporal spread between tasks and between the initial and respiking performed in Task 2A, it is also possible that there could have been systematic changes in laboratory conditions between the time periods. Many of the results from Laboratory 1 for the analytes listed in Table 7-4 that did not meet the LCMRL MQO were associated with preparation blanks with results exceeding those observed during Task 2. This may indicate that there was a higher level of background contamination in the laboratory during Task 3. This observation may indicate that the two-week period targeted in the study was not sufficient to encompass the necessary temporal variability. However, this would also be different from the results of the temporal variability assessment performed using the Task 1 blanks that are presented in Section 4.

In summary, the fact that the LCMRL MQO failure rate differed significantly from 1% could have been due to temporal changes between the initial and respiking periods, either by the difference in temporal spread between the initial and respiking data or due to changes in laboratory conditions between the initial and respiking phases. While the LCMRL procedure itself does not require a specific number of days or batches over which to spread analyses, the observed failure rates could indicate that the timeframe should, at the very least, be consistent across initial spiking and any necessary respiking.

7.2 FACDQ QL Assessment

Unlike the LCMRL, the FACDQ QL has multiple target MQOs, including an RSD, mean recovery, and false negative rate MQO. Therefore, EPA performed multiple assessments to evaluate whether the procedure accurately estimates the lowest concentration to meet the QL MQOs. To avoid biasing these assessments, it was necessary to determine which MQO was limiting, i.e., which of the three MQOs would be achieved at the highest concentration for each analyte/lab set.

Assessments of the Method 625 FACDQ QL are presented in Section 7.2.1, and assessments of the Method 200.7 FACDQ QL_{DLT} and QL_{DLK} are presented in Section 7.2.2. All comparisons were performed at the 95% confidence level. Statistical assessments of the MQOs were not performed separately for each laboratory unless otherwise noted, due to the lower statistical power.

7.2.1 FACDQ QL Assessment – Method 625

Before assessing the FACDQ QLs using Task 3 data, EPA used the Task 2 data to determine, where possible, which of the three FACDQ QL MQOs for Method 625 was achieved at the highest concentration for each analyte/lab set. This was done by examining the FACDQ QL calculations to identify any cases where either the mean recovery or the RSD MQOs failed at the initial QL spike level(s), or whether the QL was adjusted based on the FACDQ QL LER check. The former would indicate that either the recovery or the RSD MQO was limiting; the latter would indicate that the false negative rate MQO was limiting because the LER adjustment is included in the FACDQ procedure as an adjustment for meeting the false negative requirement.

If the FACDQ QL calculations did not identify a single limiting MQO, the Task 2 LCMRL data also were assessed to test whether one MQO was met at a higher concentration than the others.

Based on this assessment, EPA concluded that the 30% RSD MQO was limiting for half (72 of 144) of the analyte/lab sets for which a FACDQ QL was determined. This assessment also indicated that the 5% false negative rate was limiting for 18 of the remaining sets; these were sets for which the LER adjustment was made to the final QL. The 40-160% mean recovery MQO was determined to be limiting for only 4 analyte/lab sets. For the remaining 50 sets, no limiting MQO could be identified.

Because the choice of which MQO was limiting was not always clear, the statistical comparisons of the QLs to their target MQOs were performed twice. One comparison included only those lab/analyte sets for which that MQO was identified as limiting or for which no MQO could be identified as limiting (referred to as the "limiting MQO subgroup" in the following sections). The second comparison included all lab/analyte sets, regardless of which MQO was identified as limiting (referred to as the "full group" in the following sections).

For each analyte/lab set, only valid results for each of the seven replicates at the QL assessment level were included in the statistical assessments. The results of these assessments are presented in Sections 7.2.1.1- 7.2.1.3.

7.2.1.1 False Negative Rate Assessment

To assess the FACDQ QL false negative rate, EPA compared each of the seven replicates at the QL spike level to the FACDQ DL calculated for that analyte/lab during Task 2. The number of results falling below the DL was counted across labs and analytes for the limiting MQO subgroup and for the full group. Based on the procedure's target MQO, this percentage is expected to be 5%. To test this, EPA compared the observed frequency of false negatives to the target frequency based on a Binomial test run at the 95% confidence level. The observed frequencies and test results are presented in Table 7-5. The p-values presented in the table indicate that the observed frequency differed significantly from 5% at the 95% confidence level.

Table 7-5. FACDQ QL False Negative Rates – Method 625

Analyte/Lab Sets Included	Number of Calculated QLs	Number of Evaluated Replicates	% Replicate Results below DL	p-Value of Binomial test
Limiting MQO Subgroup	68	476	2.5%	0.0028
Full Group	144	1,008	2.3%	<0.0001

As indicated in Table 7-5, the false negative rate across all laboratories was significantly less than 5% for both the limiting MQO subgroup and the full group. Generally, the false negative rates were close to 5% for Laboratory 6 (5.7% across all sets), but were much lower than 5% for Laboratories 4 and 5 (0.6% for each laboratory across all sets).

7.2.1.2 RSD Assessment

To assess the FACDQ QL precision MQO, EPA calculated the RSD based on the seven replicates at the QL spike level for each analyte/lab set. EPA then summarized the calculated

RSDs across those labs and analytes for the limiting MQO subgroup and the full group. The distribution of the calculated RSDs is presented in Table 7-6 and Figure 7-1.

Table 7-6. FACDQ QL RSDs – Method 625

Analyte/Lab Sets Included	Number of Calculated QLs	Mean RSD	Median RSD	% RSDs exceeding 30%
Limiting MQO Subgroup	122	16.4%	13.6%	7.4%
Full Group	144	16.1%	13.4%	7.6%

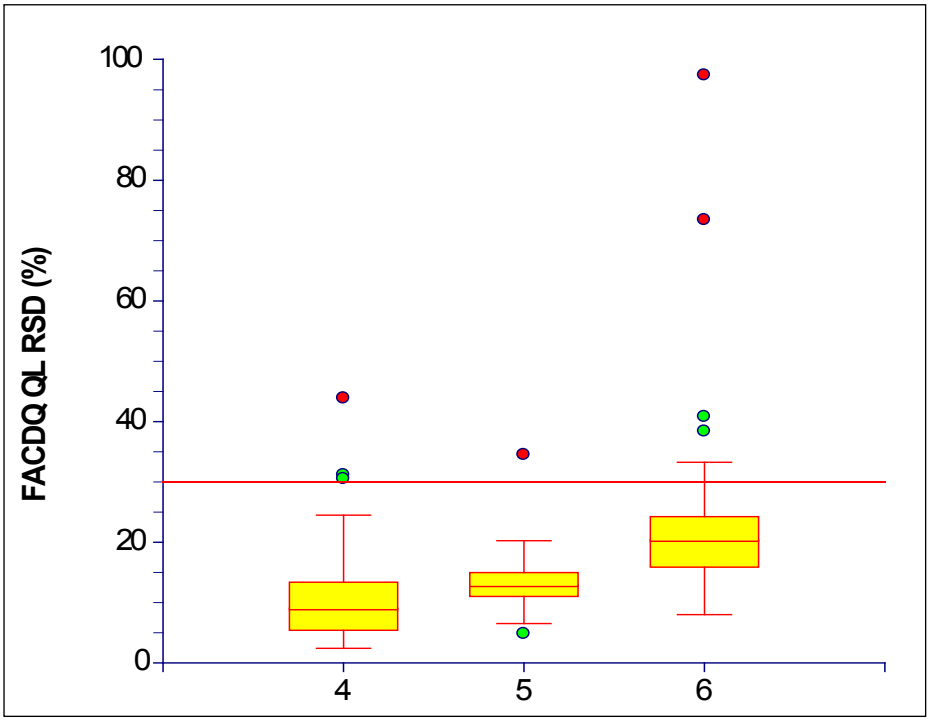

Figure 7-1. Distribution of FACDQ QL RSDs by Laboratory – Method 625

As can be seen, the distribution of calculated RSDs did not vary widely between the limiting MQO subgroup and the full group of analyte/lab sets. Generally, the RSDs tended to be closer to the 30% RSD MQO for Laboratory 6 (mean RSD of 23.4% across all lab/analyte sets) than for Laboratories 4 and 5 (mean RSDs of 11.6% and 13.4%, respectively, across all lab/analyte sets).

To assess whether the QL accurately estimates the minimum concentration at which the 30% RSD is met, EPA compared the calculated RSDs to the 30% MQO using a one-sample t-test run at the 95% confidence level. Because the D'Agostino omnibus normality test indicated that the log-transformed RSDs followed a normal distribution, the test was performed using log-transformed data. Based on this test, the hypothesis that the mean RSD at the QL equaled 30% was rejected both using the "limiting" sets ($p<0.0001$) and all sets of data ($p<0.0001$).

7.2.1.3 Mean Recovery Assessment

To assess the FACDQ QL recovery MQO, EPA calculated the mean percent bias, expressed as the absolute difference between the mean recovery and 100%, based on the valid replicates at the QL assessment level for each analyte/lab set. EPA then summarized the calculated mean percent

biases across all labs and analytes for the limiting MQO subgroup and for the full group. The distribution of the calculated mean percent biases is presented in Table 7-7 and Figure 7-2.

Table 7-7. FACDQ QL Mean Percent Biases – Method 625

Analyte/Lab Sets Included	Number of Calculated QLs	Mean Percent Bias	Median Percent Bias	% Percent Biases exceeding 60%
Limiting MQO Subgroup	54	17.3%	13.0%	1.9%
Full Group	144	21.5%	19.3%	0.7%

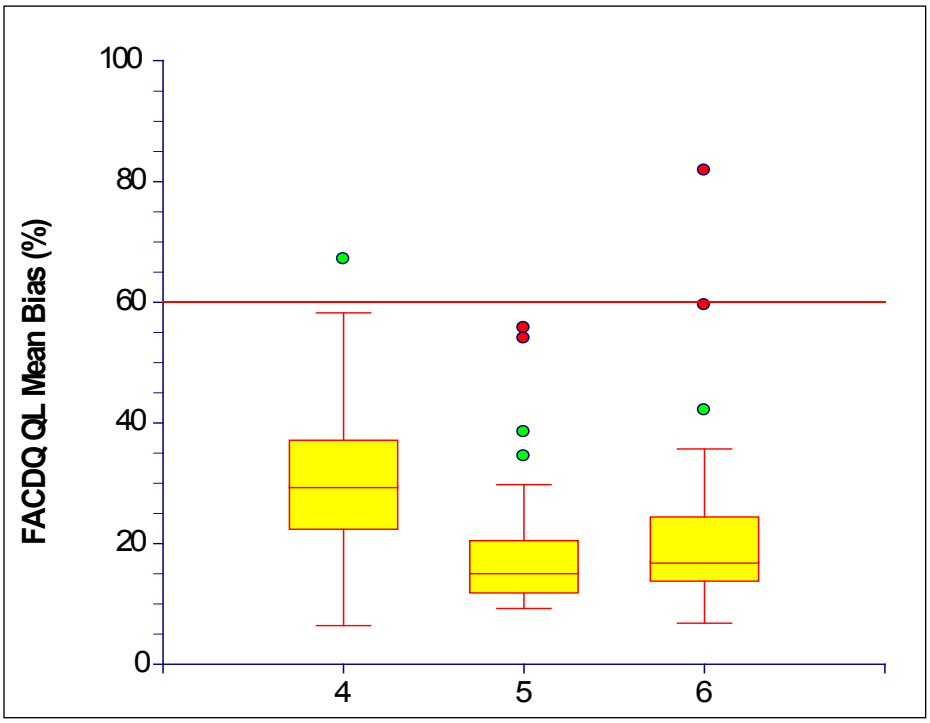

Figure 7-2. Distribution of FACDQ QL Mean Percent Biases by Laboratory – Method 625

The 60% bias MQO was met for nearly every analyte/lab set in both the limiting MQO subgroup and across the full group. Generally, the mean percent bias for Laboratory 4 (mean bias=31.6% for the full group) was closer to the MQO than for the other laboratories (mean biases of 16.6% and 16.4% for Labs 5 and 6, respectively, for the full group). As with the RSDs, EPA performed one-sample t-tests at the 95% confidence level to assess whether the QL accurately estimates the minimum concentration at which 40-160% mean recovery is met. Because the D'Agostino omnibus normality test indicated that the log-transformed mean percent biases followed a normal distribution, the test was performed using log-transformed data. Based on this test, the hypothesis that the mean percent bias at the QL equals 60% was rejected for both the limiting MQO subgroup ($p<0.0001$) and the full group ($p<0.0001$).

7.2.2 FACDQ QL_{DLT} and QL_{DLK} Assessments – Method 200.7

Two different FACDQ QLs (QL_{DLT} and QL_{DLK}) were determined for each Method 200.7 analyte/lab set. EPA followed the same approach that was used for Method 625 QLs to identify the limiting MQO for each of the two Method 200.7 QLs. Although the precision and recovery MQOs were the same for both QL_{DLT} and QL_{DLK}; each QL corresponded to a different DL (i.e.,

the QL_{DLT} with DL_T and the QL_{DLK} with DL_K). As a result, the MQO that was determined to be limiting for a given analyte/lab set was not always the same for QL_{DLT} and QL_{DLK}. Unlike Method 625, if the Task 2 data for Method 200.7 indicated that the precision and accuracy MQOs could be met at less than 2x the DL, the false negative rate was considered to be limiting for that QL even if no LER adjustment was made.

Based on this assessment, EPA determined that the 5% false negative rate was the limiting MQO for QL_{DLT} for approximately half (36 of 71) of the analyte/lab sets, and was the limiting factor for QL_{DLK} for over 80% (58 of 71) of the analyte/lab sets. For QL_{DLT}, the limiting MQO for the remaining sets was either the 20% RSD (16 sets) or could not be determined (18 sets). For QL_{DLK}, the limiting MQO for the remaining sets could not be determined for 10 of the 13 remaining sets, with the 70-130% mean recovery and the 20% RSD MQOs being determined as limiting for 2 and 1 sets, respectively.

EPA followed the same approach that was used for Method 625 to assess the Method 200.7 FACDQ QL MQOs. Separate assessments were performed for QL_{DLT} and QL_{DLK}, using data at the corresponding QL assessment levels. Similar to Method 625, statistical assessments were performed for each MQO using only those sets for which that MQO was determined to be limiting or could not be determined (the limiting MQO group), and across all analyte/lab sets (the full group). Results of these assessments are presented in Sections 7.2.2.1- 7.2.2.3.

7.2.2.1 False Negative Rate Assessment

To assess the FACDQ QL_{DLT} and QL_{DLK} false negative rates, EPA compared each of the valid replicates at the QL_{DLT} or QL_{DLK} assessment level to the corresponding FACDQ DL calculated for that analyte/lab during Task 2. The number of results falling below that DL was counted across labs and analytes for the limiting MQO subgroup and for the full group. Similar to Method 625, EPA compared the observed frequencies of false negatives to the target 5% frequency based on the Binomial test. The observed frequencies and test results are presented in Table 7-8. The p-values presented in the table indicate that the observed frequency differed significantly from 5% at the 95% confidence level.

Table 7-8. FACDQ QL False Negative Rates – Method 200.7

QL	Analyte/Lab Sets Included	Number of Calculated QLs	Number of Evaluated Replicates	% Replicate Results below DL	p-Value of Binomial test
QL_{DLT}	Limiting MQO Subgroup	54	378	0	<0.0001
	Full Group	71	497	0	<0.0001
QL_{DLK}	Limiting MQO Subgroup	68	476	0	<0.0001
	Full Group	71	497	0	<0.0001

All replicates from samples spiked at the appropriate QL level exceeded the corresponding DLs for both FACDQ QLs in both the limiting MQO subgroup and the full group. As a result, the observed false negative rates were 0%. Though this MQO was determined to be the limiting MQO for most analyte/lab sets for both QL limits, it is likely that the FACDQ procedure requirement that the QL be at least 2x the DL provided additional protection against false negatives. The lack of false negatives also could be the result of the variability of the Task 3 spiked sample results being less than that observed in the Task 1 blanks. However, this would not be consistent with the results of the temporal variability assessments presented in Section 4.

7.2.2.2 RSD Assessment

To assess the FACDQ QL precision MQO, EPA calculated the RSD based on the valid replicates at the QL_{DLT} and QL_{DLK} assessment levels for each analyte/lab set. EPA then summarized the calculated RSDs across the limiting MQO subgroup and across all labs and analytes. The distribution of the calculated RSDs is presented in Table 7-9 and Figure 7-3.

Table 7-9. FACDQ QL RSDs – Method 200.7

QL	Analyte/Lab Sets Included	Number of Calculated QLs	Mean RSD	Median RSD	% RSDs exceeding 20%
QL_{DLT}	Limiting MQO Subgroup	34	16.2%	9.1%	14.7%
	Full Group	71	15.5%	8.2%	18.3%
QL_{DLK}	Limiting MQO Subgroup	11	10.1%	8.0%	9.1%
	Full Group	71	10.1%	7.6%	11.3%

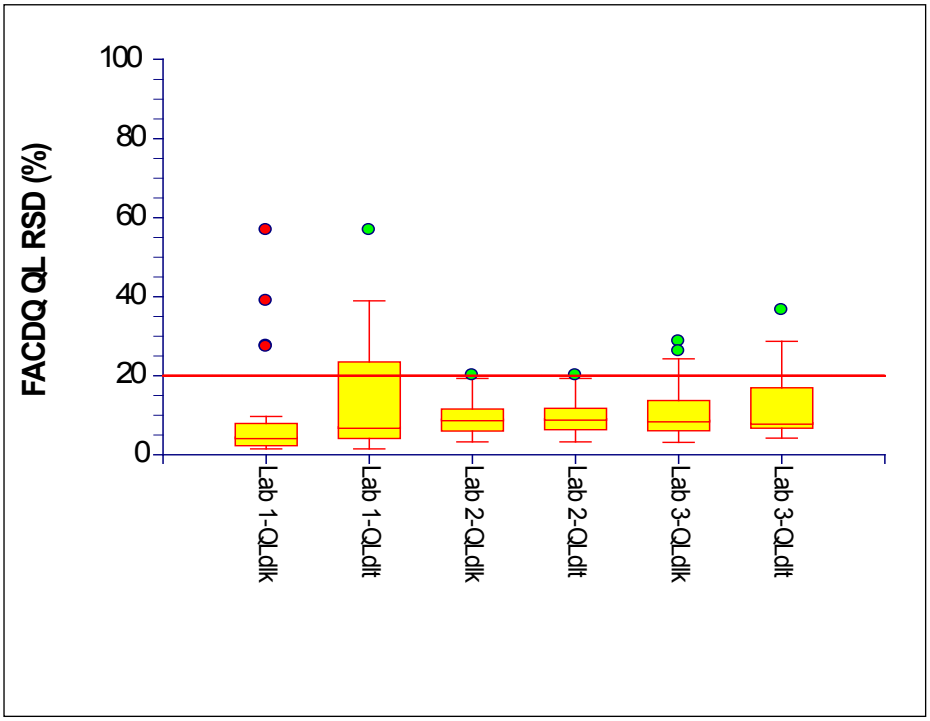

Figure 7-3. Distribution of FACDQ QL_{DLT} and QL_{DLK} RSDs by Laboratory – Method 200.7

Generally, the distribution of RSDs did not differ widely between the limiting MQO subgroup and the full group. However, the RSDs tended to be lower on average for QL_{DLK} than for QL_{DLT}. The RSDs at the QL_{DLT} assessment level tended to be closer to the 20% MQO for Laboratories 1 (mean RSD=18.0%) and 3 (mean RSD=18.7%) than for Laboratory 2 (mean RSD=9.9%); however the Laboratory 3 mean was heavily skewed due to a very high RSD for one analyte. At the QL_{DLK} assessment level, the mean RSDs tended to be similar across laboratories (mean RSDs ranging between 9.3% and 11.3%).

To assess whether the QL accurately estimates the minimum concentration at which the 30% RSD is met, EPA compared the calculated RSDs to the 30% MQO using a one-sample t-test. Because the D'Agostino omnibus normality test indicated that the log-transformed RSDs followed a normal distribution, the test was performed using log-transformed data. Based on this test, the hypothesis that the mean RSD at the QL equals 30% was rejected for both the limiting MQO subgroup and the full group for both QLs ($p<0.0001$ for all comparisons).

7.2.2.3 Mean Recovery Assessment

To assess the FACDQ QL recovery MQO, the mean percent bias, calculated as the absolute difference between the mean recovery and 100%, was calculated based on the valid replicates at the QL_{DLT} and QL_{DLK} assessment levels for each analyte/lab set. EPA then summarized the calculated mean percent biases across the limiting MQO subgroup, and across all labs and analytes. The distribution of the calculated mean percent biases is presented in Table 7-10 and Figure 7-5.

Table 7-10. FACDQ QL Mean Percent Biases – Method 200.7

QL	Analyte/Lab Sets Included	Number of Calculated QLs	Mean Percent Bias	Median Percent Bias	% Percent Biases exceeding 30%
QL_{DLT}	Limiting MQO Subgroup	19	29.5%	8.3%	5.3%
	Full Group	71	17.8%	7.0%	5.6%
QL_{DLK}	Limiting MQO Subgroup	12	6.8%	5.2%	0%
	Full Group	71	10.5%	5.8%	1.4%

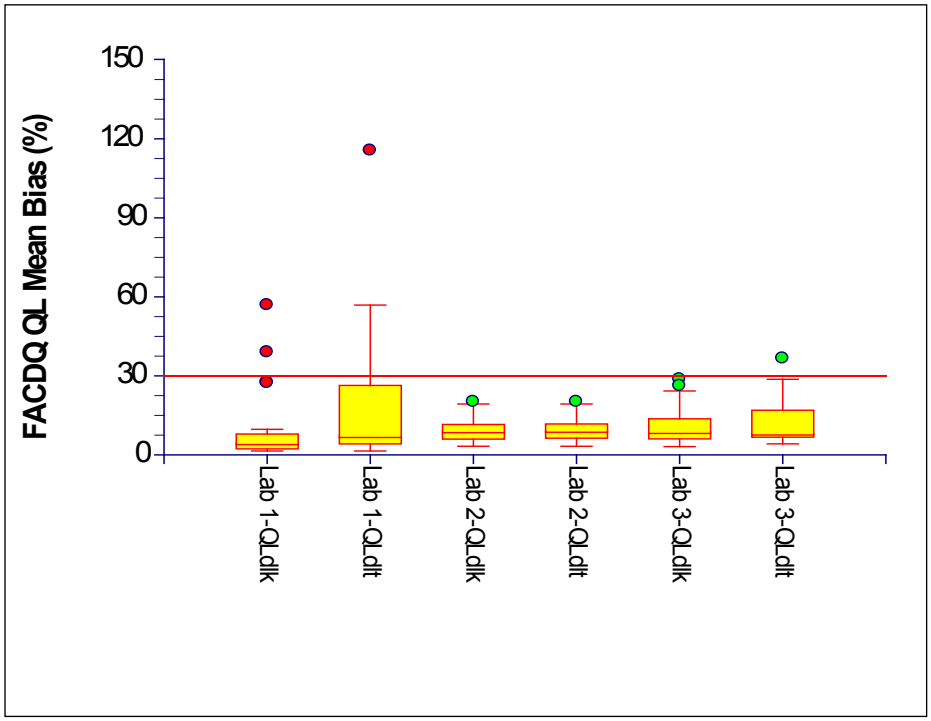

Figure 7-4. Distribution of FACDQ QL_{DLT} and QL_{DLK} Mean Percent Biases by Laboratory – Method 200.7

The mean percent biases at QL_{DLT} tended to be larger than those at QL_{DLK}. However, the median biases for the two QLs tended to be quite similar. The difference in means was due to a larger skewness observed among percent biases at the QL_{DLT} assessment level. For example, the mean percent bias for QL_{DLT} among the sets for which the recovery MQO was determined to be limiting was 29.5%, however this was due to a mean bias of 416% (i.e., mean recovery of 516%) for one analyte/lab set. No other sets for which the recovery MQO was categorized as limiting had percent bias exceeding 30% at the QL_{DLT} assessment level. The mean percent biases for both QLs tended to be higher for Laboratories 1 (QL_{DLT} mean percent bias=19.2%; QL_{DLK} mean percent bias=14.0%) and 3 (QL_{DLT} mean percent bias=30.8%; QL_{DLK} mean percent bias=13.1%) than for Laboratory 2 (QL_{DLT} mean percent bias=3.9%; QL_{DLK} mean percent bias=4.5%).

To assess whether the QL accurately estimates the minimum concentration at which the 20% mean percent bias is met, EPA compared the calculated biases to 20% using a one-sample t-test. Because the D'Agostino omnibus normality test indicated that the log-transformed percent biases followed a normal distribution, the test was performed using log-transformed data. Based on this test, the hypothesis that the mean percent bias at the QL equals 30% was rejected for both the limiting MQO subgroup (QL_{DLT}: $p=0.0006$; QL_{DLK}: $p=0.0005$) and the full group of analyte/lab sets ($p<0.0001$ for both QLs).

7.2.3 Assessment of FACDQ QL MQO Deviations

7.2.3.1 Effect of Downspiking/QL Spike Level Choice

As discussed in Section 5.2.3, some of the laboratories deviated from the downspiking steps in the FACDQ 2.4T procedure, and as a result may not have chosen appropriately low spike levels when determining their FACDQ QLs. If this was the case, the results of the MQO evaluations presented in the prior section could have been caused by this deviation rather than the FACDQ procedure not accurately estimating minimum QLs. Because these laboratories also spiked at concentrations below their determined QLs for a subset of the analytes during Task 3, EPA was able to use these data to assess whether the results of the MQO assessments would have been different if the laboratories had followed the downspiking steps more closely.

For Method 200.7, EPA reassessed the FACDQ MQOs using the valid replicates from lower spike levels that met the FACDQ downspiking requirements. As discussed in Section 5, Laboratory 3 did not performing the downspiking analyses, but instead spiked at 2x the DL during Task 2, and therefore, was assumed to have spiked at an appropriate concentration for all analytes. For Laboratories 1 and 2, EPA considered a lower spike level to be applicable if that concentration was below the QL but at or above 2x the DL, and no respiking had been required for that analyte/lab set during Task 2B due to an MQO failure. A lower spike level was not chosen for analyte/lab sets that failed one or more MQOs during Task 2, because this indicated that a spike level chosen based on the downspiking criteria would have required respiking and therefore the lab's chosen spike level for that analyte ultimately was appropriate. Based on these criteria, a new spike level was chosen for 32 analyte/lab sets among Laboratories 1 and 2 for QL_{DLT}, and for 22 of the 48 sets among Laboratories 1 and 2 for QL_{DLK}. For the remaining analyte/sets for the three laboratories, EPA used Task 3 results from the original QL assessment levels in this assessment.

Similar to Method 200.7, EPA reassessed the Method 625 FACDQ MQOs using the valid replicates from lower spike levels that met the FACDQ downspiking requirements. As discussed in Section 5.2.3, Laboratories 4 and 6 appeared to have spiked at appropriate initial QL spike levels based on their downspiking data during Task 2C For Method 625, but Laboratory 5 may not have. For Laboratory 5, EPA considered a lower spike level to be applicable if that concentration was below the Task 2 QL but at or above the spike level of the highest Task 2 downspiking sample that met all requirements (i.e., yielded an instrument signal that met qualitative identification, was within the instrument calibration range, and was at least two times the highest Task 1C blank). If respiking had been necessary for an analyte during Task 2 because either the RSD or the mean recovery MQO had not been met initially, or if the QL had been adjusted based on not meeting the LER requirement, a lower concentration was not used for this assessment. As a result, a new spike level was chosen for 40 of the 48 analytes for Laboratory 5.

Results of these assessments are presented in Tables 7-11 through 7-13. Analyses were performed following the same statistical approach described in Sections 7.2.2.1-7.2.2.3. Unlike the assessments presented in Sections 7.2.1 and 7.2.2, the distributions of RSDs and percent biases at the lower spike levels did not always follow a normal distribution, based on the D'Agostino omnibus test. Therefore, EPA used the Wilcoxon signed rank test to assess whether the lower spike level data accurately estimated the minimum concentration to meet the MQOs for those analyses for which the normality assumption was not met.

Table 7-11. FACDQ QL False Negative Rates – Lower Spike Level

Method/QL	Analyte/Lab Sets Included	Number of Calculated QLs	Number of Evaluated Replicates	% Replicate Results below DL	p-Value of Binomial test
Method 200.7 QL_{DLT}	Limiting MQO Subgroup	54	378	0%	<0.0001
	Full Group	71	497	0.4%	<0.0001
Method 200.7 QL_{DLK}	Limiting MQO Subgroup	68	476	0.2%	<0.0001
	Full Group	71	497	0.2%	<0.0001
Method 625 QL	Limiting MQO Subgroup	68	476	4.2%	0.065
	Full Group	144	1,008	3.9%	0.015

Table 7-12. FACDQ QL RSDs – Lower Spike Level

Method/QL	Analyte/Lab Sets Included	Number of Calculated QLs	Mean RSD	Median RSD	% RSDs exceeding MQO	p-Value of MQO Comparison
Method 200.7 QL_{DLT}	Limiting MQO Subgroup	34	24.2%	12.1%	29.4%	0.040
	Full Group	71	19.9%	13.3%	29.6%	<0.0001
Method 200.7 QL_{DLK}	Limiting MQO Subgroup	11	10.1%	7.9%	9.1%	0.0026
	Full Group	71	11.7%	7.9%	16.9%	<0.0001
Method 625 QL	Limiting MQO Subgroup	122	19.6%	19.2%	9.0%	<0.0001
	Full Group	144	18.9%	18.3%	9.0%	<0.0001

Table 7-13. FACDQ QL Mean Biases – Lower Spike Level

Method/QL	Analyte/Lab Sets Included	Number of Calculated QLs	Mean Percent Bias	Median Percent Bias	% Percent Biases exceeding MQO	p-Value of MQO Comparison
Method 200.7 QL_{DLT}	Limiting MQO Subgroup	19	53.5%	8.3%	10.5%	0.0059
	Full Group	71	24.6%	10.2%	11.3%	<0.0001
Method 200.7 QL_{DLK}	Limiting MQO Subgroup	12	7.6%	6.6%	0%	0.0017
	Full Group	71	12.7%	7.0%	4.2%	<0.0001
Method 625 QL	Limiting MQO Subgroup	54	13.1%	8.3%	1.9%	<0.0001
	Full Group	144	20.1%	18.3%	0.7%	<0.0001

The largest effect of the adjusted spike levels was observed in the false negative rates for Method 625, as the false negative rate did not differ significantly from 5% for the limiting MQO subgroup. The Laboratory 5 false negative rate across all analytes increased from 0.6% to 5.4%. For both methods, the RSDs and percent biases tended to increase when evaluated at the modified QL levels, but on average still were significantly lower than the target MQO. Additionally, a larger percentage of analyte/lab sets yielded RSDs and biases above the MQO target than when spiking at the QL assessment level. However, this should not be interpreted as an indication that the QLs determined in the FACDQ QL procedure cannot achieve the precision and recovery MQOs. The lower spike levels chosen for Laboratory 5 were not assessed in Task 2, and therefore, it is not known that the respiking required by the procedure when MQOs are not met would not have occurred.

7.2.3.2 Assessment of Precision/MQO Failures

As stated in the previous section, the greater number of precision and mean recovery MQO failures observed at the lower spike levels does not indicate that the procedure yields QLs at which the MQO cannot be met. However, MQO failures were observed at the original QL assessment levels for a small subset of the analyte/lab sets. In most cases, the MQO that was not met was the RSD rather than the mean recovery MQO.

As discussed in Section 7.1.2, laboratories increased the temporal spread of samples analyzed during Task 3 compared to Task 2A. This was also true for those samples that were used in the determination of FACDQ QLs (i.e., samples analyzed as part of Tasks 2B and 2C). While this would not explain the results of the MQO assessments presented in Sections 7.2.1 and 7.2.2 (i.e., that the QLs appeared to be high biased), it could explain the few unusually high RSDs observed in Task 3. While slight exceedances of the MQO would be expected to occur due to random variability if the procedure was producing accurate estimates of the minimum concentration to meet the RSD MQO, large exceedances would be unlikely. These large exceedances could indicate that the procedure may not always produce limits that meet the MQO at all.

Figures 7-5 through 7-7 show the RSDs calculated during Task 2 at the Method 200.7 QL_{DLT}, Method 200.7 FACDQ QL_{DLK}, and Method 625 FACDQ QL, respectively. In each graph, the Task 2 RSDs are compared to the RSD calculated at the corresponding QL assessment level for the same analyte and laboratory. These figures seem to show an increase in variability in Task 3

compared to Task 2 (i.e., points lying above the dashed diagonal line) for Laboratory 1 for both QLs, but little to no difference in variability for the other laboratories. As a result, it is not surprising that most of the RSD MQO exceedances observed in Task 3 for Method 200.7 were for Laboratory 1. For Method 625, most of the RSD MQO exceedances were observed for Laboratory 6. While the increase in RSDs between tasks was fairly slight, it can be seen in Figure 7-8 that the larger performance difference in Task 3 for this laboratory was a larger bias. For all analytes, this bias was negative (i.e., recoveries further below 100% in Task 3 compared to Task 2). This increased bias appeared to result in the elevated RSDs, due to the inclusion of the mean in the denominator of the RSD calculation.

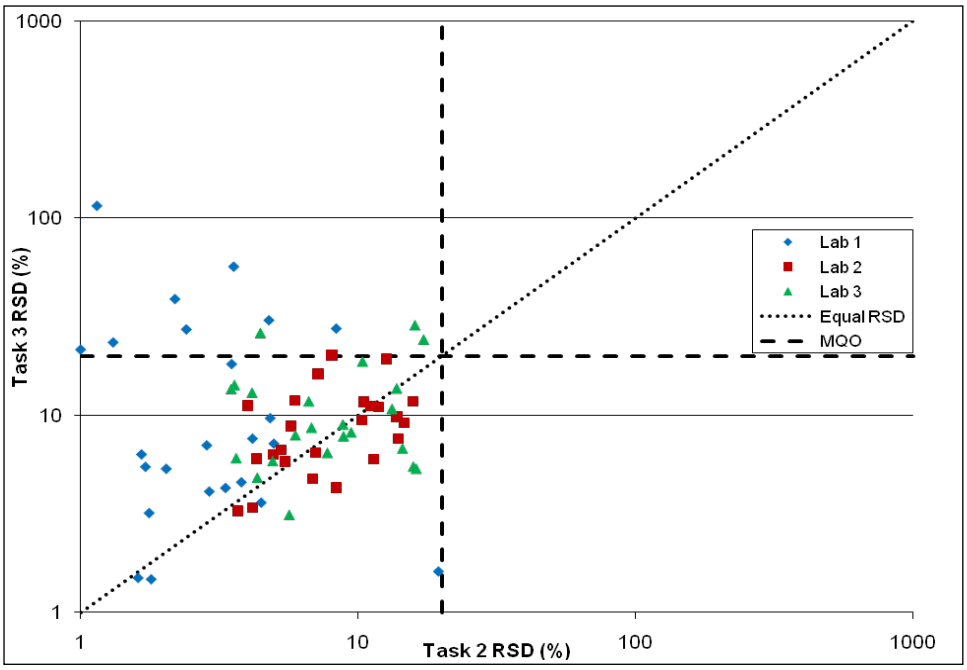

Figure 7-5. Method 200.7 Task 2 and 3 RSDs - QL_{DLT}

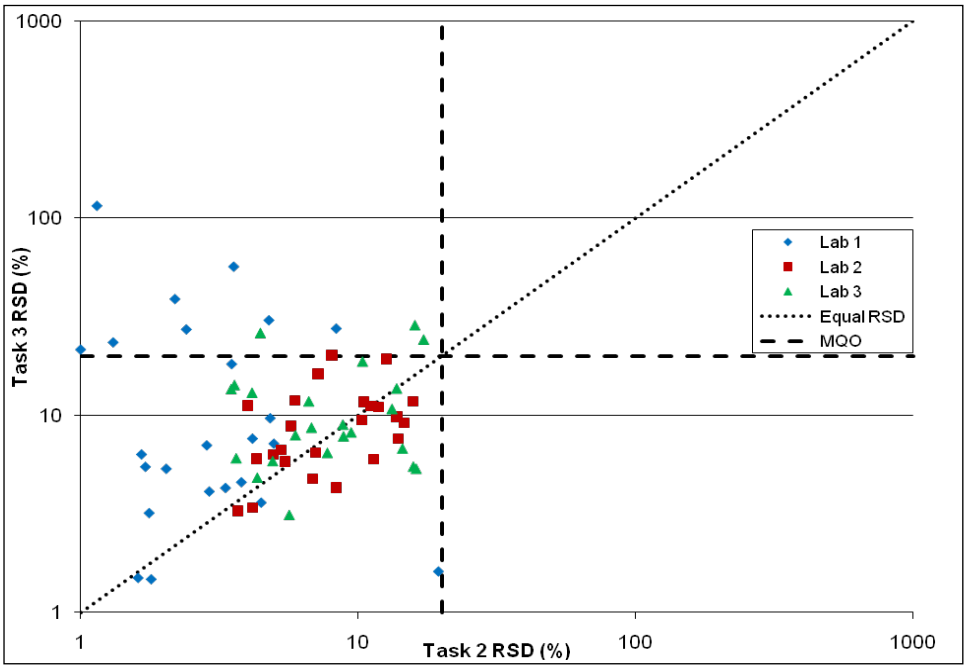

Figure 7-6. Method 200.7 Task 2 and 3 RSDs - QL_{DLK}

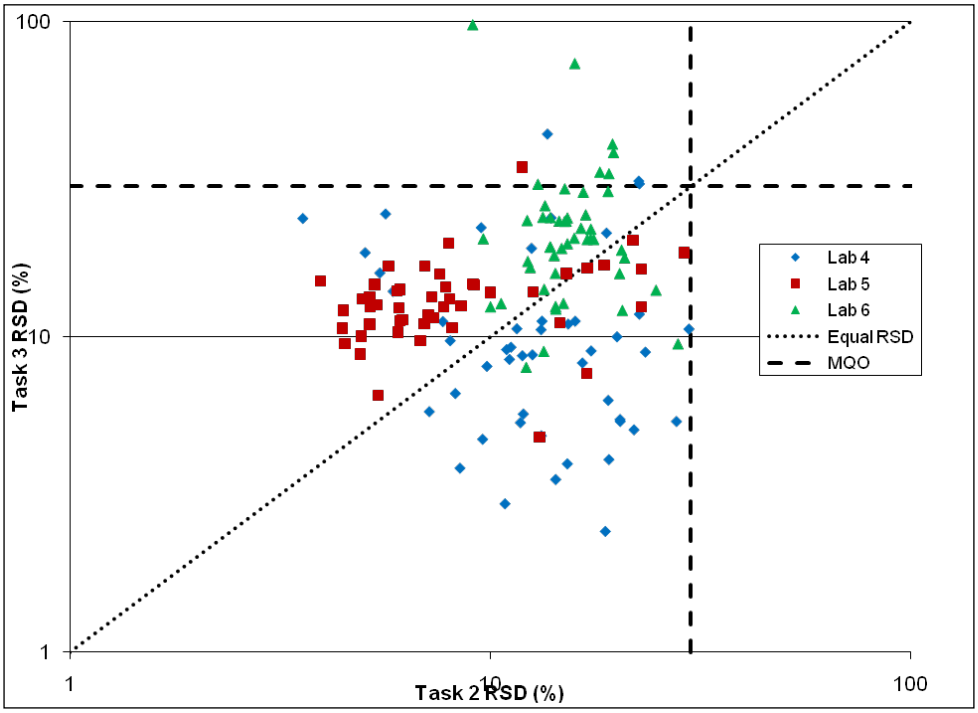

Figure 7-7. Method 625 Task 2 and 3 RSDs at the FACDQ QL

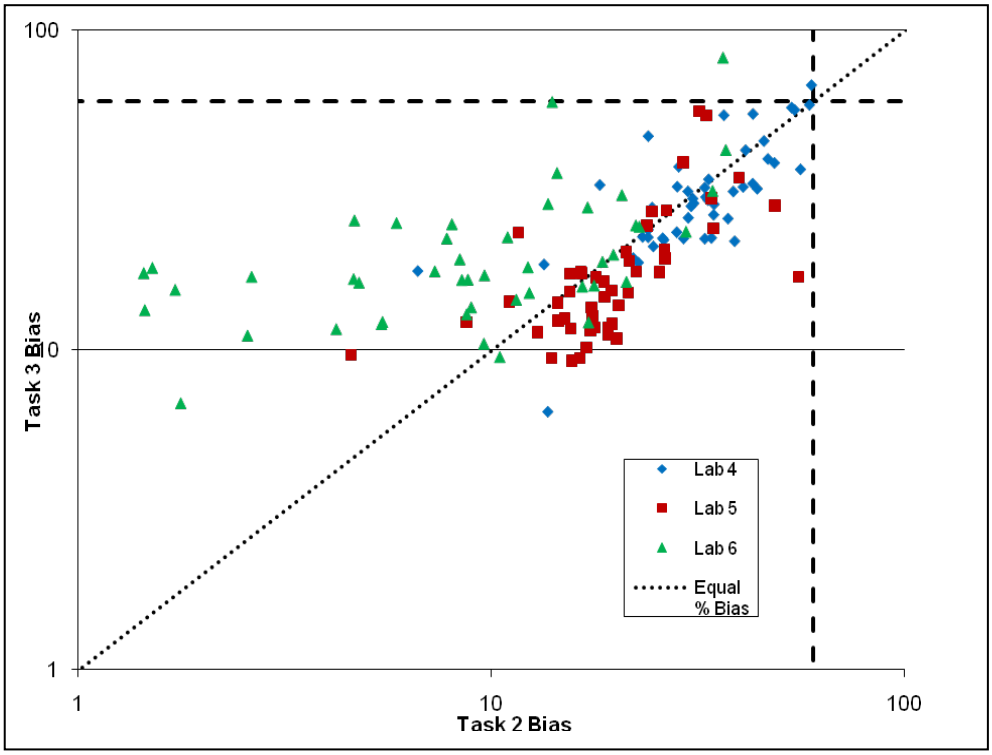

Figure 7-8. Method 625 Task 2 and 3 Mean Percent Biases at the FACDQ QL

Based on these assessments, it would appear that the MQO failures observed in the Task 3 data were largely due to changes in performance between the two tasks. Based on the instructions given in the FACDQ 2.4T procedure, QLs could be determined based on data analyzed across a year, and it is possible that MQO failures would be mitigated if this is done.

7.2.3.3 Assessment of LER Adjustment

As can be seen in Sections 7.2.1.1 and 7.2.2.1, the false negative rates for Method 625 QLs tended to be slightly higher in the limiting MQO subgroup compared to the full group of analyte/lab sets. Additionally, in Section 7.3.1.1, the false negative rate for Method 625 did not differ significantly from 5% for the limiting MQO group, while the rate did differ significantly when assessed across all analyte/lab sets. For Method 625, the false negative rate was chosen as the limiting MQO when the LER adjustment was performed on the QL. This adjustment yielded QLs that were not round numbers, and were, therefore, less feasible for spiking. QLs that did not require the LER adjustment were round numbers that had been used as spike levels during Task 2. As a result, the Task 3 QL assessment level tended to not be exactly at the QL for sets for which the false negative rate was limiting, but was exactly at the QL for sets for which one of the other MQOs was limiting.

In analyte/lab sets where the Task 3 QL assessment level was below the LER-adjusted QL, it would be expected that the false negative rate would exceed 5%. This was only observed for one set (Di-n-octyl phthalate for Laboratory 5, for which two false negatives were observed at the QL assessment level). Conversely, in cases where the QL assessment level was above the LER-

adjusted QL, it would be expected that the false negative rate would be less than 5%. In most of these cases, no false negatives were observed at the QL assessment level. While not spiking at exactly the QL does limit the ability to assess the MQOs at that QL, the false negative rate can be estimated at any concentration within the spiking range using statistical models. As stated in Section 7.2.1.1, Laboratories 4 and 5 both had low false negative rates (0.6% for each lab) at the QL assessment level. However, both of these labs analyzed and reported seven replicate results at more than the required five spike levels, and therefore, enough data were available to fit models that can be used to estimate false negative rates at the actual LER-adjusted QLs.

For each analyte for Laboratories 4 and 5, EPA calculated an observed false negative rate for each Task 3 spike level as the proportion of results that fell below that set's DL. EPA modeled these rates using a log-probit model (i.e., modeling the inverse of the standard normal cumulative distribution function based on the log-transformed spike concentration) to estimate the false negative rate for each FACDQ QL. The distribution of these false negative rates is presented in Table 7-14.

Table 7-14. Estimated False Negative Rates based on Log-Probit Model, Laboratories 4 and 5

Analyte/Lab Sets Included	Number of Calculated QLs	Mean FN Rate (%)	Median FN Rate (%)	% Percent FN Rates exceeding MQO
All	96	8.9%	4.2%	43.8%
FNR Limiting or Undetermined	48	9.5%	3.2%	37.5%
FNR Limiting only	12	25.4%	23.9%	100%

Because the false negative rates presented above are modeled estimates rather than based on observed counts, the binomial test could not be used to compare the rates to the false negative rate MQO of 5%. However, based on the nonparametric signed-rank test, these rates were not significantly different from 0 across all Laboratory 4 and 5 sets ($p=0.0513$) and for those sets in the limiting MQO subgroup ($p=0.2877$). However, the limiting MQO subgroup for the false negative rate also included sets for which no limiting MQO could be determined; the QL was not LER-adjusted for any of these sets. If the sets for which the limiting MQO was undetermined are excluded from the analysis, then the estimated false negative rates are significantly greater than the target 5% ($p=0.0005$). While this assessment is nonparametric, and as a result may be more conservative than the parametric LER adjustment included in the FACDQ procedure, it does provide some indication that the LER adjustment may not yield the appropriate false negative rate. The fitted models for two of the 12 sets for which the false negative rate was limiting are presented in Figures 7-9 and 7-10.

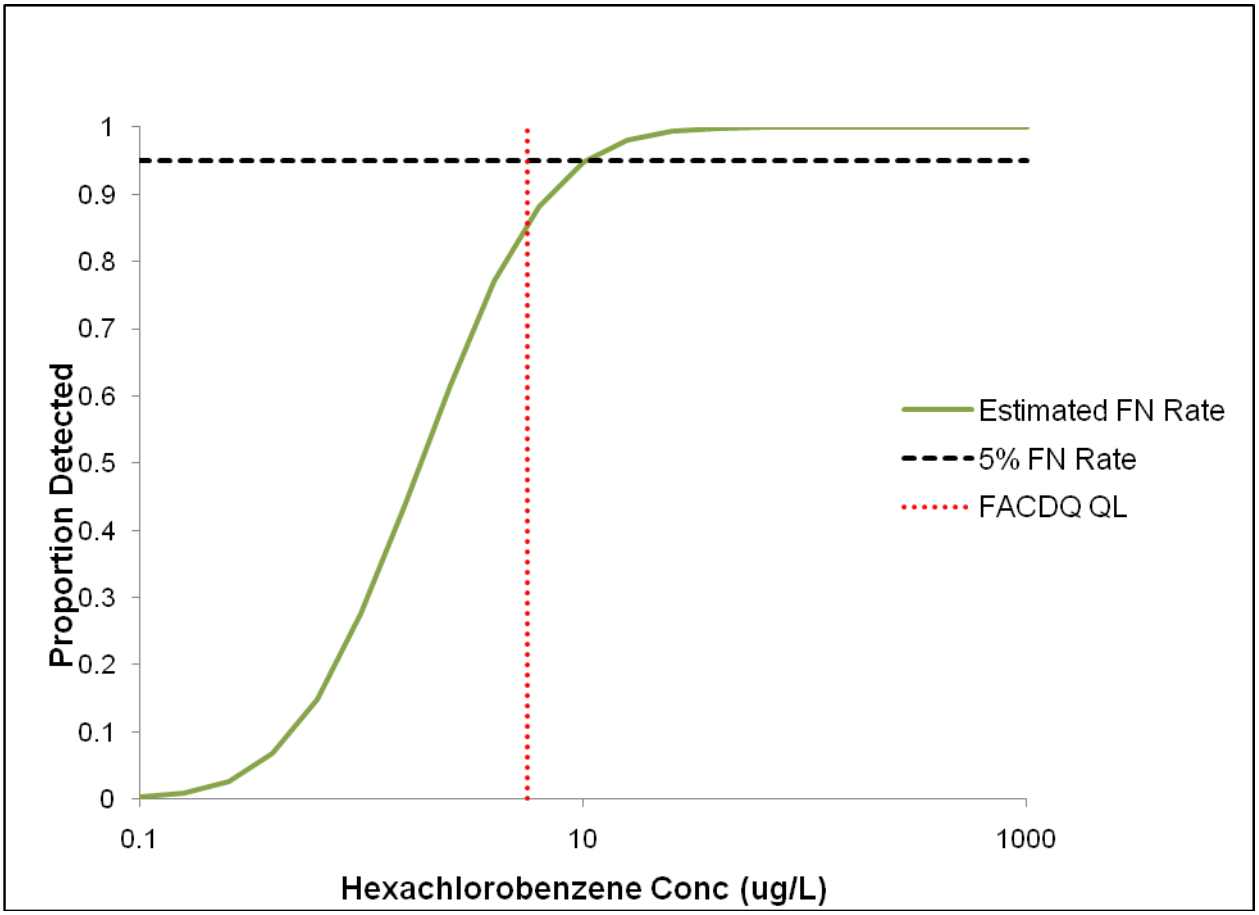

Figure 7-9. Modeled False Negative Rate – Laboratory 4 Hexachlorobenzene

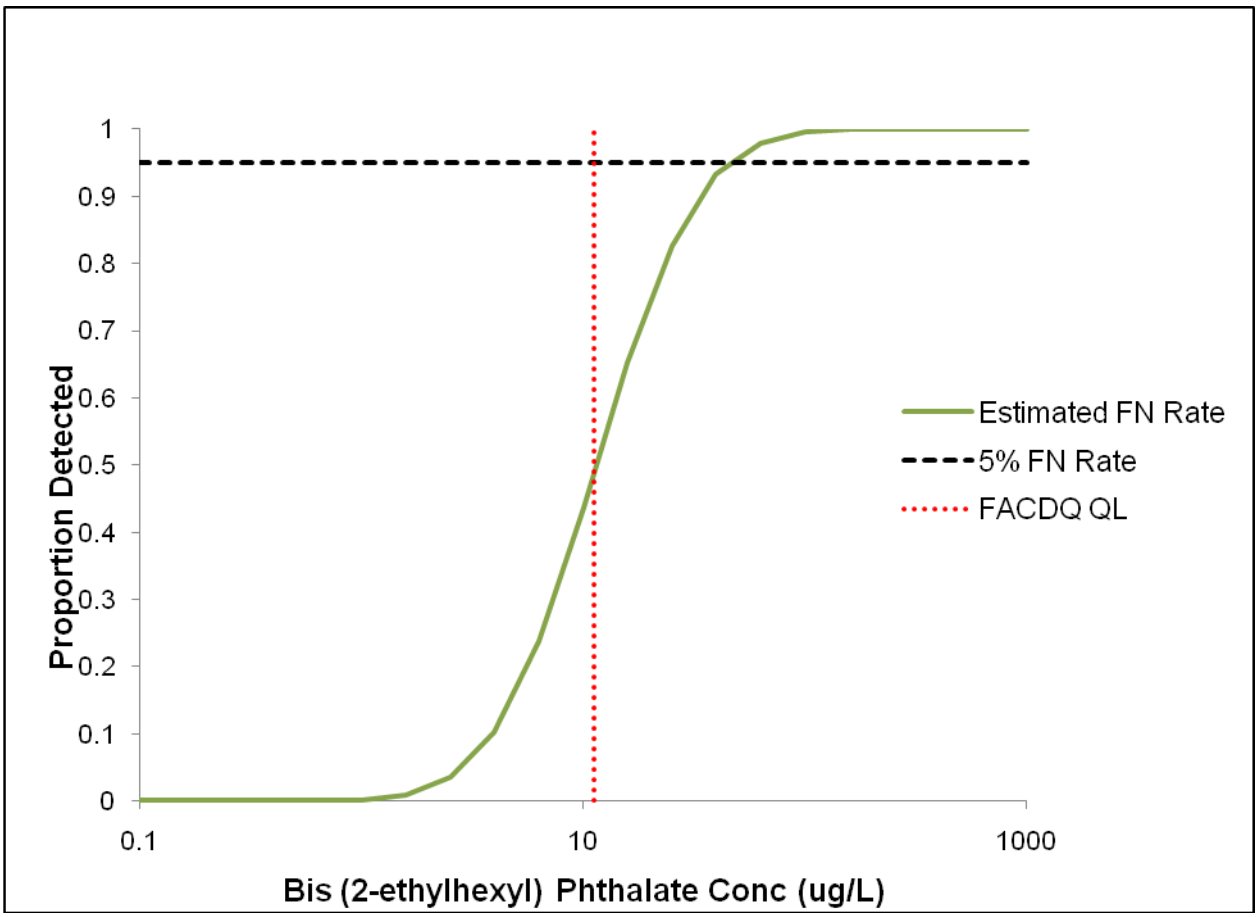

Figure 7-10. Modeled False Negative Rate – Laboratory 5 bis (2-Ethylhexyl) Phthalate

In Section 7.2.3.1, it was shown that the false negative rate for the limiting MQO subgroup was not significantly different from 5% when lower spike levels were assessed for 40 of the analytes for Laboratory 5. Of the remaining 8 analytes for which a FACDQ QL could be determined, the FACDQ QL was LER-adjusted for 4. A lower spike level was not assessed for any of these four analytes, so the results of the analysis presented in this section likely had little effect on the assessment presented in Section 7.2.3.1.

Section 8: Conclusions

The primary objective of this study was to assess whether the FACDQ procedures (versions 2.4 and 2.4T) and the LCMRL procedure can generate reliable estimates of the lowest concentration at which the procedure-specific measurement quality objectives (MQOs) can be achieved. The study results indicate that none of the procedures tested consistently generated accurate estimates of the lowest concentration at which method-specific (FACDQ procedure) or procedure-specific (LCMRL procedure) MQOs were achieved for the EPA Methods assessed in this study (Method 200.7 and Method 625). Further, none of the procedures achieved all the desired characteristics of a procedure, as defined by the FACDQ.

Specifically, for the LCMRL, Task 2 showed that after all respiking was performed, an LCMRL could be determined for 70 of the 71 analyte/lab sets for Method 200.7 and for 129 of the 147 analyte/lab sets for Method 625. Analysis of Task 3 data revealed that 7.1 % of the replicates for Method 200.7 and 2.9 % of the replicates for Method 625 did not achieve 50 to 150 % recovery (rather than the 1% frequency targeted by the procedure). Generally, when an LCMRL could not be determined for Method 625, it was because the analytical performance for that analyte was not good enough to meet the MQO, even though the laboratory spiked throughout the analytical range. This indicates that the MQO itself could not be achieved for that analyte for Method 625, rather than an issue with the LCMRL calculations. This is not surprising, given that the LCMRL was designed to be performed on drinking water samples, using analytical methods with more stringent performance expectations due to the less variable nature of drinking water samples compared to wastewater samples. It is possible that the LCMRL would have performed differently if the analytical methods used were able to achieve better performance for all analytes. Analyses of Task 3 data indicated that whether the calculated LCMRLs were the lowest concentration at which procedure-specific MQOs were achieved depended on the laboratory. The results of this assessment could have been affected by changes in laboratory performance during the duration of the study. Thus, both poor analyte performance as well as changes in laboratory performance could have affected the results for the LCMRL procedure.

An assessment of the FACDQ detection limits calculated for Method 200.7 indicated that the FACDQ DL_T tended to yield higher false positive rates than the study MQO of 1%, while the FACDQ DL_K, though it did not always meet the study MQO, tended to yield lower false positive rates than the study MQO. For both DLs, the false positive rates tended to be higher for metals more prone to background contamination. It is worth noting that the accuracy of the calculated DL may be affected because only seven blanks were used to simulate a start up DL. For Method 625, the FACDQ DL tended to yield lower false positive rates than the study MQO. This was likely due to the large number of organic analytes that are never detected in blanks which is often the case for GC/MS methods.

For the FACDQ procedure, a FACDQ QL could not be calculated for 3 out of 147 analyte/lab sets. Among those sets for which a QL could be determined, analysis of Task 3 Method 625 data indicated that:

- the false negative rate was 2.3% for replicate samples,
- 7.6% of analyte/lab sets exceeded the RSD MQO, and
- 0.7% of analyte/lab sets exceeded the recovery criteria.

Analysis of Task 3 Method 200.7 data indicated that:

- the false negative rate for both QL_{DLT} and QL_{DLK} was 0% for replicate samples,
- 18.3% of analyte/lab sets for QL_{DLT} exceeded the RSD MQO,
- 11.3% of analyte/lab sets for QL_{DLK} exceeded the RSD MQO,
- 5.6% of analyte/lab sets for QL_{DLT} exceeded the recovery MQO, and
- 1.4% of analyte/lab sets for QL_{DLK} exceeded the recovery MQO.

For most analyte/lab sets for which the recovery MQO could not be met, the RSD MQO also could not be met.

Analysis of Task 1 data indicated that false positive rates were:

- 3.42% for DL_T for Method 200.7,
- 1.13% for DL_K for Method 200.7, and
- 0.53% for Method 625.

While it cannot be determined from the results of this study, MQOs might have been more frequently met and the FACDQ QL could have been calculated for those 3 sets if the lowest possible QLs had not been targeted, and if ongoing verification, which was intended to 'self-correct' FACDQ QLs, had been performed as written in the procedure. Ultimately, it was beyond the scope and resources for this study.

Analyses of Task 3 data also indicated that the calculated FACDQ QLs were not always the lowest QL at which method/procedure-specific MQOs were achieved. This is not entirely unexpected, because FACDQ QLs targeted the lowest concentration at which multiple MQO criteria could be achieved simultaneously. This also may have been due to the spiking requirements of the procedure. The FACDQ QL was required to be at least 2 times the DL for Method 200.7, and within the calibration range for Method 625; for some analytes it appeared that the MQOs could be met below these levels. This was especially true for DL_K for Method 200.7, as mean recoveries and RSDs at QL_{DLK} were further from their MQO targets than mean recoveries and RSDs at QL_{DLT}. Additionally, the choice of spike levels used to determine the FACDQ QL in this study was heavily influenced by practical considerations of preparing and spiking at different levels for each of the large number of analytes included in the methods in this study. The performance of the procedure would likely be better for a single-analyte method; however this was beyond the scope and resources of this study.

For both the LCMRL and FACDQ QL procedures, the results of this study indicate that setting the same MQO for all analytes within a method may limit the ability of the procedures to determine accurate estimates of the lowest possible quantitation limits at which MQOs can be met. The wide differences in performance between analytes in Method 625 resulted in LCMRLs not being calculable for some analytes; this might have been mitigated if target MQOs used for those poorer performing analytes differed from the target MQOs used for the better performing analytes. For the FACDQ procedure, use of varying MQOs would minimize the likelihood that the minimum QL would occur below the procedure's allowable spiking range for some analytes, while being at a much higher level for other analytes. Additionally, it would reduce the range of spike levels across analytes when determining QLs, which could simplify the effort required to target the minimum QL.

In summary, the results of this study demonstrate that QLs calculated during this study using the FACDQ 2.4, FACDQ 2.4T and LCMRL procedures were not always the lowest QL at which method-specific (FACDQ) or procedure-specific (LCMRL) MQOs were achieved. Therefore, in light of the results of this pilot study, EPA has concluded that additional data generated using other analytical methods and more laboratories are needed to fully assess the applicability of these procedures to Clean Water Act Programs.

References

1. Report of the Federal Advisory Committee on Detection and Quantitation Approaches and Uses in Clean Water Act Programs. Submitted to the US Environmental Protection Agency December 2007. http://water.epa.gov/scitech/methods/cwa/det/upload/final-report-200712.pdf

2. Data Quality Indicators (DQIs) include aspects of method or laboratory performance, such as precision, bias, representativeness, completeness, comparability, and sensitivity. For more information see Guidance on Systematic Planning Using the Data Quality Objectives Process EPA QA/G-4. United States Environmental Protection Agency. Office of Environmental Information Washington, DC 20460. EPA/240/B-06/001 February 2006.

List of acronyms used in this report

CFR	Code of Federal Regulations
CWA	Clean Water Act
DL	Detection Limit
DL_K	Detection Limit from the FACDQ Procedure calculated using tolerance limit k
DL_T	Detection Limit from the FACDQ Procedure calculated using t-statistic
DQI	Data Quality Indicator
DQO	Data Quality Objective
EAD	Engineering and Analysis Division
EDD	Electronic Data Deliverable
FACDQ	Federal Advisory Committee on Detection and Quantitation
GC/MS	Gas Chromatography/Mass Spectrometry
ICP	Inductively Coupled Plasma
LCMRL	Lowest-Concentration Minimum Reporting Level
LER	Lowest Expected Result
MDL	Method Detection Limit (as defined in 40 CFR 136)
ML	Minimum Level
MQO	Measurement Quality Objective
MRL	Minimum Reporting Limit
ND	Not Detected
NPDES	National Pollutant Discharge Elimination System
OGWDW	Office of Ground Water and Drinking Water
PIR	Prediction Interval of Results
POTW	Publicly Owned Treatment Works
QA	Quality Assurance
QC	Quality Control
QL	Quantitation Limit
QL_{DLK}	Quantitation Limit from the FACDQ Procedure determined based on DL_K
QL_{DLT}	Quantitation Limit from the FACDQ Procedure determined based on DL_T
RSD	Relative Standard Deviation
SCC	Sample Control Center
SOW	Statement of Work