

# **Data Quality Objectives for PM Continuous Methods**

Prepared for

National Exposure Research Laboratory  
U.S. Environmental Protection Agency  
Research Triangle Park, NC 27711

Contract 68-D-00-206

ManTech Environmental Technology, Inc., Team  
P.O. Box 12313  
Research Triangle Park, NC 27709

*A ManTech International Company*

# Data Quality Objectives for PM Continuous Methods

*by*

Paul Mosquin, Frank McElroy, Andy Clayton, and Bob Vanderpool  
Research Triangle Institute

## ***ManTech Team***

*ManTech Environmental Technology, Inc.  
Research Triangle Park, NC 27709  
Prime Contractor*

*Research Triangle Institute  
Research Triangle Park, NC 27709  
Subcontractor*

*Submitted to*

Elizabeth Hunike, Work Assignment Manager  
National Exposure Research Laboratory  
U.S. Environmental Protection Agency  
Research Triangle Park, NC 27711

Contract 68-D-00-206

*Reviewed and approved by*

---

Frank McElroy  
Principal Investigator

---

Robert Vanderpool  
RTI Area Manager

ManTech Environmental Technology, Inc.  
P.O. Box 12313  
Research Triangle Park, NC 27709

*A ManTech International Company*

## Foreword

This technical report is the result of work performed by ManTech Environmental Technology, Inc., and Research Triangle Institute under Contract 68-D-00-206 for the Human Exposure and Atmospheric Sciences Division, National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC. It has been reviewed by ManTech Environmental Technology, Inc., and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

## Contents

<b>Section</b>	<b>Page No.</b>
Foreword .....	iii
Tables .....	v
1 Introduction .....	1
2 Current Equivalency Requirements .....	3
3 Data Quality Objectives .....	5
3.1 Computing the Gray Zone .....	5
3.2 Routine Sampler Performance Requirements .....	6
4 Current DQO and Equivalency Approaches .....	7
4.1 Measurement Error Models .....	7
4.2 Relating DQO and Equivalency Requirements .....	8
5 Common Measurement Error Model .....	11
5.1 Model and Assumptions .....	11
5.2 Estimating Model Parameters .....	11
5.3 Evaluating Equivalency .....	14
6 A Simulation Example .....	15
7 Discussion .....	17
8 References .....	19
Appendix A: R Code for Data Simulation and Model Fitting .....	A-1

## Tables

<b>Table</b>	<b>Page No.</b>
1 Basic Sample Design for Equivalency Tests of Class I, II, and III Samplers .....	3
2 DQO Gray Zone Maximum Relative Precision and Bias .....	6
3 Results from 1000 Simulations of Measurement Error in Current 1 in 6 Sampling Design .....	15

## 1. Introduction

The goals of this work assignment (WA) were to determine and characterize the statistical relationship between current EPA equivalent method qualification requirements and air quality monitoring data quality objective (DQO) requirements for sampling once every six days (1 in 6 sampling), and to extend this relationship in a compatible way to sampling frequencies of once every three days (1 in 3 sampling) and daily (1 in 1 sampling). In particular, the 1 in 1 sampling frequency corresponds to that used by the PM<sub>2.5</sub> continuous methods. As a result of efforts to understand the relationship, it was concluded that the relationship is indirect due to the dissimilar and largely incompatible assumptions about measurement error between the two systems. Therefore, we propose that a common measurement error model in both equivalency and DQO testing be adopted. Given a common model, the desired relationship between the two sets of requirements will naturally follow.

The national PM<sub>2.5</sub> monitoring network consists of a network of air-quality samplers used for determining compliance with federal 2.5-micron particulate matter standards. Compliance exists at a monitoring site if the three-year sample average<sup>1</sup> of measured PM<sub>2.5</sub> concentrations is less than 15.05 µg/m<sup>3</sup>.<sup>2</sup> A variety of quality control (QC) requirements are associated with this network, and we can classify these as follows:

1. *Equivalent method requirements.* A sampler model can be certified for use within the PM<sub>2.5</sub> network in one of two ways. The first, a non-statistical approach to certification, is that the sampler be designed according to specified engineering standards. Samplers of this kind are known as federal reference method (FRM) samplers and may be certified by EPA for use in the network upon demonstration of adherence to all FRM specifications. The second approach is statistical and is available for those samplers whose designs do not exactly comply with the FRM specifications. These samplers are categorized as Class I, II, or III depending on their divergence (increasing) from the FRM specifications, with Class III applicable to the continuous methods of special interest here. For each of these classes, there is (or will be developed) a specific equivalency test, in which a sampler that passes the test is considered a federal equivalent method (FEM) sampler and may then be used within the PM<sub>2.5</sub> network. The test itself consists of a comparison of daily readings at one or more test sites between some number of candidate samplers and FRM samplers to determine how closely the candidate sampler measurements correspond to those of the FRM samplers.
2. *Data quality objectives (DQOs).* Associated with the PM<sub>2.5</sub> network are DQOs used to evaluate network performance in two ways:
  - a. A site is in compliance if its three-year average particulate matter concentration is less than 15.05 µg/m<sup>3</sup>. The DQOs recognize that some uncertainty is associated with this decision and so provide a simulation-based statistical methodology (U.S. EPA,

---

<sup>1</sup> There is also a 24-hour standard that is of minor interest in this WA.

<sup>2</sup> Five one hundredths are added to account for rounding error.

2002) to compute a “gray zone” interval, which covers  $15.05 \mu\text{g}/\text{m}^3$ . If a three-year average lies within the interval, then the compliance decision is somewhat weak. The current DQO gray zone has been developed for 1 in 6 samplers with FRM sampler characteristics, a maximum relative bias<sup>3</sup> of 0.1, and maximum relative precision<sup>4</sup> of 0.1.

- b. The DQOs require that the performance of the network as a whole be monitored to ensure that samplers and laboratories are providing readings of adequate relative precision or relative bias. The gray zone is developed based on certain allowable ranges of relative precision and relative bias, and so a system of collocating a field sampler with a second sampler has been developed to estimate their values at a site or aggregate site level. If estimated values at a site or aggregate level are outside of allowable ranges, then further action may be taken.

This report is primarily concerned with parts 1 and 2a, aiming to first understand any relationship between the sampler performance requirements used to construct the 1 in 6 gray zone and the current equivalency requirements for 1 in 6 candidate samplers. It then aims to provide guidelines for equivalency testing of 1 in 3 and 1 in 1 sampling based on any understanding gained, as well as proposed sampler requirements for maintaining acceptable gray zones with 1 in 3 and 1 in 1 sampling. As described in this report, these guidelines require that equivalency testing methodology be changed so that its model for sampler measurement error is compatible with the one used in constructing DQO gray zones.

This paper is organized as follows: Section 2 provides a more detailed background on the equivalency testing approach. Section 3 presents the background for the DQOs, both in constructing the gray zone and in testing field samplers. Section 4 offers a statistical comparison of the DQO and equivalency testing. Section 5 describes issues in developing a common framework for the two approaches. Section 6 provides the suggested approach.

---

<sup>3</sup> The relative bias  $b$  of estimator  $\hat{\theta}$  of  $\theta$  is defined according to  $E[\hat{\theta}] = (1 + b)\theta$  where  $E[\hat{\theta}]$  is the expectation of  $\hat{\theta}$ .

<sup>4</sup> The relative precision  $\tau$  of random variable  $X$  of mean  $\mu$  and standard deviation  $\sigma$  is given by  $\tau = \frac{\sigma}{\mu}$ .

## 2. Current Equivalency Requirements

Equivalency requirements are those requirements that candidate samplers must satisfy to obtain FEM status as a result of an equivalency test against an FRM sampler. Candidate samplers are first classified as Class I, II, or III, and then, depending on the class, PM<sub>2.5</sub> collocated measurements are collected at a specific number of test sites, with a certain number of reference and candidate samplers per site, for a minimum number of days.

The sample design for each of the three classes of candidate samplers is given in Table 1. As is evident from the table, there is a reasonable amount of variation in sample design according to class. That is, the number of sites selected, number of reference samplers per site, number of candidate samplers per site, and number of measurement days per site can all vary by sampler class.

**Table 1.** Basic Sample Design for Equivalency Tests of Class I, II, and III Samplers

Class	Number of Sites	Reference Samplers per Site	Candidate Samplers per Site	PM <sub>10</sub> Samplers per Site <sup>a</sup>	Days of Measurements per Site
I	1	3	3	0	10
II	2	3	3	1	10
III	(4) <sup>b</sup>	(2)	(2)	0	(120) <sup>c</sup>

<sup>a</sup>PM<sub>10</sub> samplers are used to ensure that the Class II sampler data satisfy PM<sub>2.5</sub>/PM<sub>10</sub> ratio requirements.

<sup>b</sup>Values in parentheses are proposed.

<sup>c</sup>For continuous candidate samplers, the recorded reading is the average of 1-hour readings.

For data collected according to one of the above designs to be considered acceptable, it must satisfy the following requirements:

1. Daily site means of the reference method samplers
2. Daily site standard deviations and relative precisions for reference methods
3. Days of the year on which sampling can occur

Concerning requirement 1 for the means of the daily reference method site averages, it is required that enough daily means fall both above and below a certain value. For example, with a Class I sampler, means must fall within a range of 10–200 µg/m<sup>3</sup>, and of the required minimum total of 10 days' measurements from one site, at least three means must be below 30 µg/m<sup>3</sup> and three means above that level. This requirement ensures that means are neither all high nor all low, as might happen if a test was conducted during a period of minimal atmospheric variation.<sup>5</sup>

Concerning requirement 2 for the daily site standard deviations and relative precisions, these depend on whether a daily site reference average is above or below a specified amount. If it is above

<sup>5</sup> Current practice has been for data to be discarded if it does not satisfy these requirements. Potentially data might also be discarded if enough of a certain type was already obtained. It might be suggested that a policy of keeping all data be implemented, with outliers to be later removed on agreement of EPA.

the amount, then the relative precision must satisfy a requirement. If it is below the amount, then the standard deviation (precision) must satisfy a separate requirement. For example, with a Class I sampler, if a daily reference average exceeds  $40 \mu\text{g}/\text{m}^3$ , then the percent relative precision must be less than 5%; if the daily site average is less than  $40 \mu\text{g}/\text{m}^3$ , then the standard deviation should be less than  $2 \mu\text{g}/\text{m}^3$ . The use of standard deviation for lower valued means is due to the large variance of relative precision in these cases. In either case, this requirement might be viewed as a method for outlier detection and removal.

Finally, the temporal requirements for the (proposed) Class III methods are that at least five measurements be collected per month per site, with 30 measurements per quarter. Class III methods do not have any range requirements of variety, so these temporal requirements play that role for these methods.<sup>6</sup>

Once a data set is considered acceptable, the equivalency of the candidate and reference method can be tested. This equivalency is established if at each site statistics based on daily mean concentrations of candidate and reference methods fall within specified ranges. More specifically, if there are  $I$  sites,  $J$  samplers per site, and  $K$  measurements per sampler per site, and if

$X_{ijk}^*$  is the  $\text{PM}_{2.5}$  measurement at time  $k$  for reference sampler  $j$  at site  $i$ , and

$X_{ijk}$  is the  $\text{PM}_{2.5}$  measurement at time  $k$  for candidate sampler  $j$  at site  $i$ ,

then at site  $i$  one would compute the reference means  $\bar{X}_{ik}^*$  and the candidate means  $\bar{X}_{ik}$  at time  $k$ . The candidate sampler is then equivalent if for each site  $i = 1, \dots, I$

- a least squares or simple linear regression of candidate means on reference means has intercept estimate  $\hat{\alpha}_i$  in the interval  $-1 \leq \hat{\alpha}_i \leq 1$  and slope parameter  $\hat{\beta}_i$  in  $0.95 \leq \hat{\beta}_i \leq 1.05$ , and
- the Pearson's sample correlation coefficient  $\hat{\rho}_i$  of the correlation of the  $\bar{X}_{ik}^*$  and the  $\bar{X}_{ik}$  satisfies  $\hat{\rho}_i \geq 0.97$ .

These are the equivalency requirements for 1 in 6 samplers.

---

<sup>6</sup>With the exception of a requirement that site reference means fall in an overall range of 5–200  $\mu\text{g}/\text{m}^3$ .

### 3. Data Quality Objectives

The DQOs for the PM<sub>2.5</sub> network have two important uses: (1) to evaluate the strength of compliance decisions using a gray zone and (2) to allow continual monitoring of network performance in the field. A brief overview of each is now provided.

#### 3.1 Computing the Gray Zone

The gray zone is computed using a simulation-based approach that incorporates both population and measurement error models. The population model generates the true daily mean, while the measurement error model adjusts the daily true mean for error associated with the sampler. The current gray zone for compliance decision is the interval (12.2,18.8). A site is either in or not in compliance, depending on which side of 15.05 µg/m<sup>3</sup> its three-year site means lie, but this compliance decision is not considered strong if the three-year average also lies within the gray zone.

The population model holds that daily mean concentrations at a site follow a sinusoidal curve with an annual period of 1. Highest particulate matter concentrations occur in the summer, and the lowest in the winter. The sine curve has parameters chosen according to an analysis of observed particulate readings in the current PM<sub>2.5</sub> network. To allow for population variability, the value from the sine curve for a given day is multiplied by a log-normally distributed error of mean 1 and standard deviation 0.8, thereby providing the day's true PM<sub>2.5</sub> value.

The DQO simulation adjusts the true daily PM<sub>2.5</sub> values for measurement error. If  $\mu_k$  is the true PM<sub>2.5</sub> value at time k, the final recorded value is then equal to

$$X_k = (1 + b)\mu_k + Z_k,$$

where the  $Z_k$  are independent, normally distributed  $Z_k \sim N(0, [(1 + b)\tau\mu_k]^2)$ , with variance depending on the true daily PM<sub>2.5</sub> value, as well as the relative bias  $b$ , and relative precision  $\tau$ .<sup>7</sup> In the simulations, a relative bias of at most  $b = 0.1$  and a relative precision of at most  $\tau = 0.1$  was assumed. These values are considered reasonable performance standards for the PM<sub>2.5</sub> samplers and are the DQO (gray zone) requirements for 1 in 6 samplers.

Multiple simulations of the above model provided a gray zone interval of (12.2, 18.8). The bounds of this interval were chosen to have the following property: If the true yearly population mean particulate matter concentration equals the upper bound, then the probability of simulating a three-year average less than 15.05 µg/m<sup>3</sup> is 5%, and if the true yearly population mean equals the lower bound, then the probability of simulating a three-year average greater than 15.05 µg/m<sup>3</sup> is 5%.

Increasing the sampling frequency from 1 in 6 to 1 in 3 and 1 in 1 allowed the relative bias requirement to be relaxed and yet still maintain the same gray-zone interval. Table 2 gives the

---

<sup>7</sup> Note that the model can be expressed with an additive or multiplicative error, where the multiplicative form is given as  $X_k = (1 + b)\mu_k Z_k^*$  where  $Z_k^* = N(0, \tau^2)$ .

current and proposed relative biases and precisions based on gray-zone simulation of 1 in 6, 1 in 3, and 1 in 1 sampling.

**Table 2.** DQO Gray-Zone Maximum Relative Precision and Bias

Sampling Frequency	Relative Precision	Relative Bias
1 in 6	0.1	0.1
1 in 3	0.1	0.13
1 in 1	0.1	0.18

In this report, one of the goals is to provide compatible equivalency testing requirements for these proposed 1 in 3 and 1 in 1 gray-zone requirements.

### 3.2 Routine Sampler Performance Requirements

An additional DQO for the PM<sub>2.5</sub> network is that over time both PM<sub>2.5</sub> samplers and laboratories continue to perform at acceptable levels. Performance checks occur at both site and aggregate levels, with the performance parameters of interest being the two parameters of the DQO measurement error model: relative precision and relative bias.

To check network performance with respect to relative bias, four collocated measurements are taken at one quarter of the sites each year. The chosen sites rotate so that over four years, all sites in the network are assessed. For these collocated samplers, daily estimates of relative bias are computed and then averaged over the quarterly estimates to obtain a yearly estimate. Yearly estimates are then further aggregated and averaged up to reporting level to check for biases within a network. Due to the limited sample size per site, inferences are only made at the aggregate level as a means of checking method performance.

Relative precision of network samplers is evaluated through having a fixed one-quarter of all sites equipped with a permanent collocated sampler. Daily estimates of precision based on the paired measurements are averaged by year and/or averaged by site for reporting regions. If averages are not over time, then comparisons may be made via a control chart approach.

## 4. Current DQO and Equivalency Approaches

This section describes similarities and differences between the measurement error model used for equivalency testing and that used to obtain the DQO gray zone. This comparison of models may allow for statements to be made about what equivalency testing requirements should be for 1 in 3 and 1 in 1 sampling. However, the conclusion is that this is difficult at best, and, for this reason, it is suggested that a common measurement error model be introduced for both purposes. Since the DQO measurement error model is the more realistic model (as described below), it is suggested that the statistical approach for equivalency testing be based on this model.

The section has two main parts. The first is a description of the measurement error models in each case, and the second is a discussion of how the requirements in each case might be related.

### 4.1 Measurement Error Models

The equivalency testing approach, as described in section 2, does not have an explicit measurement error model. For a candidate sampler to obtain FEM status, sample statistics based on least squares and correlation analysis must satisfy certain range requirements. However, we can consider the least-squares approach to imply a measurement error model, namely, the simple linear model, which implies two assumptions:

1. The  $\bar{X}_{ik}^* = \frac{1}{J} \sum_j X_{ijk}^*$  reference means are measured with certainty.
2. The candidate means are independent, normally distributed  $\bar{X}_{ik} \sim N(\alpha_i + \beta_i \bar{X}_{ik}^*, \frac{\sigma^2}{J})$ , where  $\sigma^2$  is the variance of the individual sampler measurements.

The measurement error model for the DQO gray-zone simulations was given in section 3.1. Adding indices to the observations for site, sampler, and time gives

$$X_{ijk}^* = (1 + b_{ij}^*) \mu_{ik} + Z_{ijk}^*$$

$$X_{ijk} = (1 + b_{ij}) \mu_{ik} + Z_{ijk}$$

where the  $Z_{ijk}^* \sim N(0, [(1 + b_{ij}^*) \tau_{ij}^* \mu_{ik}]^2)$  and  $Z_{ijk} \sim N(0, [(1 + b_{ij}) \tau_{ij} \mu_{ik}]^2)$ , independently.

Expressing the DQO model in terms of site averages gives the following:

1. The reference means  $\bar{X}_{ik}^*$  are independent, normally distributed  $\bar{X}_{ik}^* \sim N((1 + \bar{b}_i^*) \mu_{ik}, \frac{\mu_{ik}^2}{J^2} \sum_j (1 + b_{ij}^*)^2 (\tau_{ij}^*)^2)$ , where  $\bar{b}_i^*$  is the mean of the relative biases of reference samplers at site  $i$ , and  $\tau_{ij}^*$  is the precision of sampler  $j$  at site  $i$ .

2. The candidate means  $\bar{X}_{ik}$  are independent, normally distributed

$\bar{X}_{ik} \sim N\left((1 + \bar{b}_i)\mu_{ik}, \frac{\mu_{ik}^2}{J^2} \sum_j (1 + b_{ij})^2 \tau_{ij}^2\right)$ , where  $\bar{b}_i$  is the mean of the relative bias of candidate samplers at site  $j$ , and  $\tau_{ij}$  is the precision of sampler  $j$  at site  $i$ .

It is clear that the two measurement error models are very different.<sup>8</sup> Crucial assumptions underlying the equivalency model are that measurement errors for candidate samplers are of constant variance and that reference sampler site daily means are not random. In contrast, the DQO model allows for randomness in both candidate and reference sampler measurements, and allows the variance for these measurements to depend on the mean. For samplers of the PM<sub>2.5</sub> variety, having the variance depend on the mean is considered a more realistic assumption.

Although the equivalency assumptions do not entirely hold in practice, they do lead to a simple comparison of reference to candidate samplers under 1 in 6 sampling assumptions. However, with the desire to extend equivalency comparisons to 1 in 3 and 1 in 1 sampling using the DQO requirements as a guide, the difference between the approaches becomes more important.

In making the extension from the current 1 in 6 DQO requirements to the equivalency testing requirements, as a guide we might ask how the current sets of requirements in each case relate, even though the measurement error models are different. This is the topic of the next section.

## 4.2 Relating DQO and Equivalency Requirements

Finding relationships between the DQO and equivalency requirements is not simple due to the differing underlying parameterizations and measurement error models. However, since a project goal is to outline existing relationships, in this section we will attempt to do so. The first step is to ask how a given parameterization in one model might be interpreted in terms of the other. Given an interpretation, we might then ask how the current requirements relate. Since the proposed 1 in 3 and 1 in 1 DQO requirements are to be used to find corresponding equivalency requirements, we begin by considering how the DQO requirements might be interpreted for equivalency testing.

The DQO relative bias parameter is represented within the equivalency testing error model through its contribution to the slope of the regression line

---

<sup>8</sup>Later in this paper, the simplifying assumptions that reference sampler relative biases equal zero and that relative precisions among a given sampler type are equal (i.e.,  $\tau_{ij} = \tau$  and  $\tau_{ij}^* = \tau^*$  for all  $i, j$ ) are made. Under these assumptions, the DQO site sample means have the following:

1. The reference means  $\bar{X}_{ik}^*$  are independent, normally distributed  $\bar{X}_{ik}^* = N\left(\mu_{ik}, \frac{(\mu_{ik}, \tau^*)^2}{J}\right)$ .
2. The candidate means  $\bar{X}_{ik}$  are independent, normally distributed  $\bar{X}_{ik} \approx N\left((1 + \bar{b}_i)\mu_{ik}, \frac{\mu_{ik}^2}{J^2} \sum_j (1 + b_{ij})^2 \tau^2\right)$ .

$$\beta_i = \frac{\sum_j (1 + b_{ij})}{\sum_j (1 + b_{ij}^*)}$$

That is, the slope at site *i* is the ratio of the average of the multiplicative biases for candidate and reference samplers. The equivalency test requires that the estimate of this ratio fall between 0.95 and 1.05. Since the DQO model requires a sampler relative bias of no more than 10%, it follows that if only one sampler of each kind were compared,<sup>9</sup> the regression slope would allowably fall within (.81,1.22) since the minimum allowable is 0.9/1.1=.81 and the maximum is 1.1/0.9=1.22. These bounds remain the same as additional samplers are added to the sites (provided there are equal numbers of both kinds). If the additional assumption is made of reference samplers having zero relative bias, then the ratio is simply the average of multiplicative biases for the candidate samplers. In this case, the DQO minimum for one sampler is 0.9 and the maximum is 1.1, which is already outside of the equivalency bounds. The equivalency standards for relative bias appear to be more conservative than those required for the DQO gray zone.

In contrast to relative bias, there does not appear to be a statistic within the equivalency tests that either directly or indirectly estimates the DQO relative precision. Neither the sample correlation coefficient nor the estimated regression intercept contributes this information. Also, the daily site observations are reduced to a daily mean, and the data are screened so that daily site level reference sample variances are no more than a given amount. Both of these steps remove some of the sampler-specific variation.

The equivalency test sample correlation coefficient is a function of the estimated slope and the sums of squares for reference and candidate samplers, that is,

$$\hat{\rho}_i = \frac{S_{\bar{X}\bar{X}^*}}{\sqrt{S_{\bar{X}\bar{X}}}\sqrt{S_{\bar{X}^*\bar{X}^*}}} = \hat{\beta}_i \frac{\sqrt{S_{\bar{X}\bar{X}}}}{\sqrt{S_{\bar{X}^*\bar{X}^*}}}$$

where  $S_{\bar{X}\bar{X}}$ ,  $S_{\bar{X}^*\bar{X}^*}$ , and  $S_{\bar{X}\bar{X}^*}$  are the sums of squares and cross products for the daily site means. The ratio of sums of squares is equal to the ratio of sample variances, and so the slope and this ratio can be considered instead of the slope and the correlation coefficient. The requirement that the sample correlation coefficient be greater than or equal to 0.97 is therefore also a joint requirement on the fitted slope and the ratio of sample variances of the candidate and reference means. These sample variances estimate population variability of PM<sub>2.5</sub> measurements for those days included in the sample, however, and not sampler precision.

The estimated regression intercept of the equivalency test shifts the fitted regression line from the origin and allows for contamination within the reference and candidate samplers. As such, it might be viewed as the average of the sum of the contaminations for both groups of samplers. For

---

<sup>9</sup>Note that current equivalency requirements do not depend on either the number of samplers in the comparison or the number of observations collected.

example, if the candidate samplers consistently have a contamination of  $1 \mu\text{g}/\text{m}^3$ , while the reference samplers have zero, we might expect the regression intercept to equal 1.

It appears, therefore, that the above comparisons provide little guidance for extending current requirements in a compatible way for 1 in 3 and 1 in 1 sampling. The relationships appear to be indirect, at best, and, furthermore, since the DQO and equivalency requirements have historically developed independently, there remains the question of the extent to which any existing relationship should matter. Since the proposed 1 in 3 and 1 in 1 DQO requirements (see Table 2) are to be used to develop compatible 1 in 3 and 1 in 1 equivalency requirements, we suggest instead changing the equivalency testing measurement error model to be the same as that used in the DQO process. Doing so will make the extension clear and direct.

## 5. Common Measurement Error Model

The difficulties encountered in section 4 suggest that a common measurement error model be used for both the DQO gray zone and the equivalency testing. This model will be based on the current DQO gray-zone model, which has more realistic assumptions about the measurement error process.

### 5.1 Model and Assumptions

From section 4.1, the DQO gray-zone measurement error model as applied to equivalency testing can be written as

$$X_{ijk}^* = (1 + b_{ij}^*)\mu_{ik} + Z_{ijk}^*$$

$$X_{ijk} = (1 + b_{ij})\mu_{ik} + Z_{ijk}$$

where  $Z_{ijk}^* \sim N(0, [(1 + b_{ij}^*)\tau_{ij}^*\mu_{ik}]^2)$ , and  $Z_{ijk} \sim N(0, [(1 + b_{ij})\tau_{ij}\mu_{ik}]^2)$ . Assumptions are that

- both  $X_{ijk}^*$  and  $X_{ijk}$  are measured with error;
- error standard deviation increases multiplicatively with the mean;
- bias is multiplicative;
- errors are independent and normally distributed; and
- relative bias and precision are sampler specific.

However, in practice, it is reasonable to assume that candidate and reference samplers have common relative precisions, that is,  $\tau_{ij}^* = \tau^*$  and  $\tau_{ij} = \tau$  for all  $i, j$ . It is also reasonable to assume that reference samplers have zero relative bias, that is,  $b_{ij}^* = 0$  for all  $i, j$ . The zero relative bias assumption will statistically identify the  $b_{ij}$  and is in keeping with the observed performance of reference samplers.

### 5.2 Estimating Model Parameters

The DQO measurement error model is a multiplicative error model. That is, the additive form of the model given earlier can be written as

$$\begin{aligned} X_{ijk}^* &= \mu_{ik} Z_{ijk}^* \\ X_{ijk} &= (1 + b_{ij}) \mu_{ik} Z_{ijk} \end{aligned}$$

where in this case  $Z_{ijk}^* \sim N(1, (\tau^*)^2)$  and  $Z_{ijk} \sim N(1, \tau^2)$ , independently. This assumes zero relative bias for the candidate samplers and common precisions for candidate and reference samplers, as described earlier.

Since the  $Z_{ijk}^*$  and  $Z_{ijk}$  are normal random variables, there is a small but positive probability that they might take negative values, in which case  $X_{ijk}^*$  and/or  $X_{ijk}$  would lie outside of their observable range. Furthermore, it is likely that  $\tau^* \leq 0.1$ , and,  $\tau \leq 0.1$ , in which case the normal distribution for the errors will be closely approximated by a lognormal distribution. That is,  $Z_{ijk}^* \approx LN(1, (\tau^*)^2)$  and  $Z_{ijk} \approx LN(1, \tau^2)$ .

Logarithms can now be taken, and a reparameterization done:

$$Y_{ijk}^* = \log X_{ijk}^* \tag{1.1}$$

$$\begin{aligned} &= \log \mu_{ik} + \log Z_{ijk}^* \\ &= \theta_{ik} + \varepsilon_{ijk}^* \end{aligned}$$

$$Y_{ijk} = \log X_{ijk} \tag{1.2}$$

$$\begin{aligned} &= \log \mu_{ik} + \log(1 + b_{ij}) + \log Z_{ijk} \\ &= \theta_{ik} + \gamma_{ij} + \varepsilon_{ijk} \end{aligned}$$

where  $\varepsilon_{ijk}^* \approx N(0, \sigma_{N^*}^2)$  and  $\varepsilon_{ijk} \approx N(0, \sigma_N^2)$ , independently. The model can now be fit using weighted (or generalized) least-squares, although to do so requires estimates for the  $\sigma_{N^*}^2$  and  $\sigma_N^2$  to obtain the weights.

The estimation of  $\sigma_{N^*}^2$  is relatively straightforward, as there are replicate reference observations at each site and time. Since the reference samplers are assumed unbiased, site level reference sample variances  $S_{N^*,ik}^2$  can be estimated for each  $i,k$ , and then an unbiased estimate of  $\sigma_{N^*}^2$  obtained as their average:

$$\hat{\sigma}_{N^*}^2 = \frac{1}{IK} \sum_{i,k} S_{N^*,ik}^2$$

The  $S_{N^*,ik}^2$  are independent, identically distributed (by model assumption), and

$$\frac{(J-1)IK \sum_{i,k} S_{N^*,ik}^2}{\sigma_{N^*}^2} \square \chi_{(J-1)IK}^2$$

is distributed as a chi-square  $\chi^2_{(J-1)IK}$  random variable with  $(J-1)IK$  degrees of freedom. This distribution can be used to construct confidence intervals or hypothesis tests for  $\sigma_{N^*}^2$ .

The estimation of the candidate error variance  $\sigma_N^2$  must be done differently, as unlike for the reference observations, the replicate candidate measurements at each site estimate different means due to the candidate-specific relative biases. To estimate  $\sigma_N^2$ , an estimate of  $\theta_{ik}$  based solely on the reference samplers is used. This estimate is obtained as the average of the reference sampler readings at a given site and time  $\hat{\theta}_{ik}^* = \frac{1}{J} \sum_j Y_{ijk}^*$ . The variance of the estimated difference  $Y_{ijk} - \hat{\theta}_{ik}^*$  is equal to

$$\text{Var}_j(Y_{ijk} - \hat{\theta}_{ik}^*) = \sigma_N^2 + \frac{\sigma_{N^*}^2}{J}$$

since the reference and candidate observations are independent. For sampler  $j$  at site  $i$ , this variance can be unbiasedly estimated using the sample variance of the differences, and a sampler-specific estimate  $\hat{\sigma}_{N,ij}^2$  of the candidate variance then be obtained as

$$\hat{\sigma}_{N,ij}^2 = \text{Var}_j(Y_{ijk} - \hat{\theta}_{ik}^*) - \frac{\hat{\sigma}_{N^*}^2}{J}.$$

Since this is a difference in variance estimates, there is the potential for a negative estimate, particularly in small samples or when the candidate samplers are more precise than the reference samplers.

The error variance of the candidate samplers is then estimated as the mean of the variance estimates for individual candidate samplers

$$\hat{\sigma}_N^2 = \frac{1}{IJ} \sum_{ij} \hat{\sigma}_{N,ij}^2$$

This estimate is unbiased, although not independent of the variance estimate for the reference samplers.

Using the above variance estimates, weighted least squares can be performed with weights taken as the reciprocals of the estimated variances. An example using simulated data is provided in the next section.

For final estimates  $\sigma_{N^*}^2$ ,  $\sigma_N^2$ ,  $\hat{\gamma}_{ij}$ , a back transformation to the original parameter space is accomplished as follows:

$$\hat{\tau}^* = \sqrt{e^{\hat{\sigma}_{N^*}^2} (e^{\hat{\sigma}_{N^*}^2} - 1)}$$

$$\hat{\tau} = \sqrt{e^{\hat{\sigma}_N^2} (e^{\hat{\sigma}_N^2} - 1)}$$

$$\hat{b}_{ij} = e^{\hat{\gamma}_{ij}} - 1$$

Note that if as expected  $\tau^* \leq 0.1$  and  $\tau \leq 0.1$  then  $\tau^* \approx \sigma_{N^*}$ , and  $\tau \approx \sigma_N$ , and the above might be written with a single variance parameterization.

### 5.3 Evaluating Equivalency

Having a common measurement error model for DQO and equivalency testing greatly simplifies the equivalency testing, since the desired performance levels for candidate samplers (see Table 2) can now be estimated directly after completion of the test.

Because of the simplifying model assumption that sampler-specific precisions are equal for all candidate samplers, a common candidate sampler relative precision parameter is estimated from the model. The estimated candidate relative precision can then be compared directly to the value required by the DQO gray zone.

For relative bias, however, each candidate sampler has its own sampler-specific relative bias estimated, and so there is the question of how to decide whether the candidate samplers as a whole are satisfying the DQO requirements. One approach would be to consider the candidate samplers equivalent if all estimated biases satisfied the DQO requirement. Another approach would be to assume a normal distribution for the bias parameters and consider the candidate samplers equivalent if the estimated probability of a randomly selected candidate sampler with a relative bias outside the requirements is less than a given amount.

To assume that the relative bias parameters have a normal distribution would be to treat their estimates as random effects. Ideally, the model would explicitly do that (currently they are treated as fixed effects), although deciding if the normal assumption is reasonable would require more candidate samplers in the tests. The model might be made a random-effects model from the start, although to do so would require time to develop such an approach.

Ideally all estimates should be evaluated with an associated level of uncertainty, either using confidence intervals or hypothesis tests. This is a limitation of the current approach, although it might be addressed with additional work on the modeling.

## 6. A Simulation Example

A simulation using the current 1 in 6 equivalency tests—1 site, 3 samplers per type, 10 daily measurements—is now provided to illustrate the performance of the model.

The underlying population data for the simulation was obtained using the DQO sinusoidal model. Ten days were randomly selected from the year, and the true  $PM_{2.5}$  means for those days were multiplied by lognormal variables of mean 1 and standard deviation 0.8 to adjust for population variability. These population values were accepted as is, without further screening. For the three candidate samplers, relative bias parameters were drawn from an assumed  $N(.05, .03^2)$  distribution. The mean and variance of this distribution lead to candidate samplers with a tendency to bias readings upwards. The population and relative bias values were then retained for 1000 simulations of measurement error.

Table 3 illustrates the unbiasedness of the simulation estimates for the log-transformed model and provides some insight on the precision of the estimates. Notably, there were at times negative variance estimates for the candidate sampler precision, due in part to the sample size, and also to the candidate sampler having been specified as more precise than the reference sampler for this simulation.

**Table 3.** Results from 1000 Simulations of Measurement Error in Current 1 in 6 Sampling Design

Original Data			Log-Transformed Data				
Parameter	True Value	Simulation Mean	Parameter	True Value	Simulation Mean	Simulation Standard Error	Simulation Standard Deviation
$\tau^*$	.05	.049	$\sigma_{N^*}$	.05	.050	$2.5 \times 10^{-4}$	.008
$\tau$	.025	.025	$\sigma_N$	.025	.025 <sup>1</sup>	NA	NA
$b_{11}$	.075	.076	$\gamma_{11}$	.073	.073	$3.8 \times 10^{-4}$	.012
$b_{12}$	.038	.039	$\gamma_{12}$	.038	.038	$3.8 \times 10^{-4}$	.012
$b_{13}$	.097	.098	$\gamma_{13}$	.093	.093	$3.8 \times 10^{-4}$	.012

<sup>a</sup>Obtained as the root of the average variance, as opposed to the average of the root of simulation variances, due to the occasional negative variance estimate. Standard errors and deviations are not provided for this reason. When a negative variance estimate was encountered, the candidate precision was set equal to one-half of the estimated reference precision for purposes of obtaining weights. Doing so might increase standard deviations of the estimated relative biases, although the estimates themselves are still unbiased.

If we assume that the bias parameters are normally distributed, then we might ask what the probability would be of a candidate sampler having relative bias outside of the DQO required bounds using the estimated biases to estimate the parameters of the normal distribution. With only three estimates of relative bias, this provides an admittedly rough approach. To illustrate, according to the true normal distribution for the relative bias parameters, this probability would be 0.048. If the relative biases of the three candidate samplers of the simulation example were known with certainty, then the mean of the normal distribution would be estimated as 0.070 and its standard

deviation as 0.030. The probability of observing a relative bias outside of the acceptable range is then 0.16. If the simulation estimate relative biases were used to estimate this probability, then the average estimate of the probability would be 0.15 with a standard deviation of 0.09.

Code in the R language is provided in Appendix A both for simulating data under the measurement error model and for fitting the model reading data from a file.

## 7. Discussion

In this report, a comparison of the current 1 in 6 equivalency and DQO requirements has been provided, as well as a methodology to conduct the equivalency tests using the same measurement error model as that used in the DQO modeling. We now discuss general aspects of equivalency testing.

Although it is not the focus of this report to redesign the equivalency tests, many additional questions might be asked. Should range requirements be used? If so, what should they be? How many candidate or reference samplers should be used? How many sites should be used? How many days should be used? These questions might, in part, be addressed by simulation using the given model.

The method provided in this report estimates precision parameters for both candidate and equivalency samplers. With enough observations, the precision of reference samplers may be known well enough to use a constant value. Doing so might improve the precision estimates for the candidate samplers. However, estimating it separately does provide a check of sorts on the results.

Other checks on the model assumptions might be accomplished through comparison of the  $\hat{S}_{Nij}^2$  or the  $\hat{\sigma}_{N^*,ij}^2$ . Graphical visualization of the model fit might be done by plotting the site means of candidate versus reference samplers, or by using the estimated daily means (the  $\hat{\mu}_{ik}$ ) to create plots of the individual observations.

Various scenarios might be investigated using these approaches. For example, there may be a temperature effect on sampler bias where readings in one time period might be biased in an opposite direction to those in another. Such a situation would not affect the estimates of relative precision, but would lead to relative bias estimates lying between the two extremes. If effects such as these are thought to exist, then the model might be extended to accommodate them, perhaps by adding separate seasonal bias parameters, or by allowing the bias to depend on temperature in the model.

If checks suggest a problem with the model, then it should be noted that the adopted model is essentially the DQO model (with the exception of assumptions of section 5.1), and so failure of the model might then suggest modification of both the equivalency and DQO measurement error models. If needed, the model should allow extension to models with predictors such as temperature and humidity.

If the model does require extension, one potential direction is suggested in the paper by Rocke and Lorenzato (1995). In their paper they considered models of the form

$$X = \alpha + \beta\mu Z + W,$$

where the  $\beta\mu Z$  gives the current multiplicative model, with  $\alpha$  now allowing for contamination and  $W$  being an additive error allowing for error at low measured particulate matter concentrations. Their model would need modification to allow for separate error distributions of candidate and reference samplers and for sample-specific  $\beta$ 's, and consequences for model fitting might either be an increase in the dimension of integration (if random effects) or an increase in the dimension of the space being optimized over (if fixed effects).

An alternative approach to extending the model would be to place it within a Bayesian framework. Doing so would lead to the ability to fit the model in a single step (as opposed to first estimating variances to obtain weights) and would provide probability statements about parameters of interest that would account for both statistical and practical significance. As with all Bayesian analyses, prior distributions would be required for the model parameters; however, given vague priors and sufficient data, their effect on the final inference would be negligible.

## 8. References

Rocke, D.M., and Lorenzato, S.L. 1995. A two-component model for measurement error in analytical chemistry. *Technometrics* 37(2):176–184.

U.S. EPA. 2002. DQO Companion, Version 1.0 User's Guide. EPA Work Assignment 5-07.

## Appendix A

### R Code for Data Simulation and Model Fitting

This appendix provides the software code to simulate observations under the population and measurement error model and also to load observed (or simulated) values from a file to fit the weighted least-squares model.

Simulations and model fitting for this WA were conducted in the R language, a high-quality, freely available data analysis package based on the S language. The software can be found at [www.r-project.org](http://www.r-project.org) for most major operating systems. The commercially available S-Plus language is also based on S, and the code given here should run in S-plus with minor modification. The code can be run either by cut and pasting into the R command prompt window or through use of the source command to load directly from a file.

Three sections of code are provided. The first section gives generic functions that should be loaded into R prior to using either of the other two sections. The second section provides code to simulate observations and save to a file. The third section provides code to read observations from a file and fit the model.

An example of the input file format is given in Figure A-1. The file is tab-delimited, with the first column giving the site and the second the sampler at that site. Remaining columns are measurements from time 1 through time K. A header row must provide variable names, although these are not used in the code. The candidate and reference sampler data should be stored in separate files with this format.

Site	Sampler	t1	t2	t3	t4
1	1	4.316	21.775	49.893	6.967
1	2	4.498	21.072	47.852	6.014
1	3	4.251	19.682	49.464	6.901
2	1	0.973	6.569	7.2	20.666
2	2	0.869	6.338	7.016	19.099
2	3	0.88	5.812	6.62	19.367

**Figure A-1.** Example of file format for candidate or reference sampler data.

## Generic R Functions

```

# Function to add measurement error to daily means
add.merr <- function(muik,sdev,rb){

  zijk <- rnorm(I*J*K,mean=1,sd=sdev)
  zijk <- array(zijk,dim=c(I,J,K))

  xijk <- array(0,dim=c(I,J,K))

  for(i in 1:I){
    for(j in 1:J){
      for(k in 1:K){
        xijk[i,j,k] <- muik[i,k]*(1+rbcij[i,j])*zijk[i,j,k]
      }
    }
  }
  return(xijk)
}

# get reference variance, log transformed data
get.varlogr <- function(yijk){
  s2ik <- array(0,dim=c(I,K)) # Estimate reference variance
  for(i in 1:I){
    for(k in 1:K){
      s2ik[i,k] <- var(yijk[i,,k])
    }
  }
  return(mean(s2ik))
}

# get candidate variance, log transformed data
get.varlogc <- function(yijk,xijk,var1){
  sigma2ij <- array(0,dim=c(I,J)) # Estimate candidate variance
  est.daymean <- rep(0,K)
  for(i in 1:I){
    for(k in 1:K){
      est.daymean[k] <- log(mean(xijk[i,,k]))
    }
    for(j in 1 :J){
      sigma2ij[i,j] <- var(yijk[i,j,]-est.daymean)-var1/J
    }
  }
  return(mean(sigma2ij))
}

# Function to write PM25 readings to tab-delimited file.
write.pm25 <- function(filepath,filename,xijk){
  header <- c("site","sampler", paste("t",1:K,sep=""))
  write.table(cbind(rep(1,J),1:J,round(xijk[1,,],3)),
              paste(filepath,filename,sep=""),sep="\t",quote=F,
              row.names=F,col.names=header)
  if(I>1){

```

```

    for(i in 2:I){
      write.table(cbind(rep(i,J),1:J,round(xijk[i,,],3)),
                 paste(filepath,filename,sep=""),sep="\t",append=T,quote=F,
                 row.names=F)
    }
  }
}

# Function to read PM25 readings to tab-delimited file.
read.pm25 <- function(filepath,filename){
  x <- read.table(paste(filepath,filename,sep=""),sep="\t",header=T)
  xijk <- array(0,dim=c(I,J,K))
  for(i in 1:I){
    for(j in 1:J){
      for(k in 1:K){
        xijk[i,j,k] <- x[I*(i-1)+j,k+2]
      }
    }
  }
  return(xijk)
}

```

## R Code to Simulate Data

```

# Define constants
I <- 1          # Number of sites
J <- 3          # Samplers per site
K <- 10         # Readings per sampler per site
sdr <- .05     # Reference precision
sdc <- .025    # Candidate precision
rbc.mean <- .03 # Candidate relative bias random effects mean
rbc.sd <- .05  # Candidate relative bias random effects standard deviation
FILEOUT <- T   # Write data to file?
FILEPATH <- "C:\\temp\\" # File Path. Double backslash Windows, Single backslash UNIX.
FILENAMER <- "PM25R.dat" # Reference sampler data file
FILENAMEC <- "PM25C.dat" # Candidate sampler data file

rbcij <- array(0,dim=c(I,J))          # Generate true sampler relative biases
rbcij[1,] <- rnorm(J,mean=rbc.mean,sd=rbc.sd) # Assume normal distribution

muik <- array(0,dim=c(I,K))
days <- sample(1:365,K)
muik[1,] <- 16.3+11.125*sin((2*pi*days)/365+1.9) # Population model mean daily
x <- rnorm(K,mean=-.25,sd=.8)
muik[1,] <- muik[1,]*exp(x) # Adjust for daily variation with lognormal(1,.8^2)

xrijk <- add.merr(muik,sdr,array(0,dim=c(I,J)))
xcijk <- add.merr(muik,sdc,rbcij)

if(FILEOUT==T){ #Write to file

```

```

write.pm25(FILEPATH, FILENAMER, xrijk)
write.pm25(FILEPATH, FILENAMEC, xcijk)
}

```

## R Code to Fit Model

```

# Define constants
I <- 1      # Number of sites
J <- 3      # Samplers per site
K <- 10     # Readings per sampler per site
rblb <- -.1 # DQO relative bias lower bound
rbub <- .1  # DQO relative bias upper bound
FILEPATH <- "C:\\temp\\" # File Path. Double backslash Windows, Single backslash UNIX.
FILENAMER <- "PM25R.dat" # Reference sampler data file
FILENAMEC <- "PM25C.dat" # Candidate sampler data file

xrijk <- read.pm25(FILEPATH, FILENAMER) # Read from file
xcijk <- read.pm25(FILEPATH, FILENAMEC)

yrijk <- log(xrijk)      # log-transform
ycijk <- log(xcijk)

y <- rep(0, I*J*K*2) # Arrange regression response vector
for(i in 1:I){
  for(j in 1:J){
    for(k in 1:K){
      y[(i-1)*J*K+(j-1)*K+k] <- yrijk[i, j, k]
      y[I*J*K+(i-1)*J*K+(j-1)*K+k] <- ycijk[i, j, k]
    }
  }
}

site <- rep(rep(1:I, rep(J*K, I)), 2) # Find regression predictor vectors
sampler.cand <- matrix(0, nrow=I*J*K, ncol=I*J)
for(i in 1:(I*J)){
  sampler.cand[((i-1)*K+1):(i*K), i] <- rep(1, K)
}
sampler <- rbind(matrix(0, nrow=I*J*K, ncol=I*J), sampler.cand)
time <- rep(rep(1:K, J), I), 2)
site <- as.factor(site)
time <- as.factor(time)

varlogr <- get.varlogr(yrijk)
varlogc <- get.varlogc(ycijk, xrijk, varlogr)
varlogc2 <- varlogc
if(varlogc<=0) varlogc2 <- varlogr/2 # If negative variance estimate
# then set to one half reference variance
w <- c(rep(1/varlogr, I*J*K), rep(1/varlogc2, I*J*K)) # Weight vector

m <- lm(y~-1+sampler+site*time-time-site, weights=w) # The regression

# Standard deviations and relative biases for log transformed data

```

```
sdevlogr <- sqrt(varlogr)
if(varlogc>0) sdevlogc <- sqrt(varlogc)
rblogc <- m$coefficients[1:J]

# Standard deviations and relative biases for raw data
sdevr <- sqrt(exp(varlogr)*(exp(varlogr)-1))
if(varlogc>0) sdevc <- sqrt(exp(varlogc)*(exp(varlogc)-1))
rbc <- exp(rblogc)-1

# probability candidate outside bounds
reffects.mean <- mean(rbc)
reffects.sdev <- sqrt(var(rbc))
prob.rb <- pnorm(rblb,mean=reffects.mean,sd=reffects.sdev) +
  1-pnorm(rbub,mean=reffects.mean,sd=reffects.sdev)

# output to screen, if needed
print(paste("Estimated reference sampler precision:",round(sdevr,3)))
if(varlogc<0){
  print("Estimated candidate sampler precision: Negative variance")
} else {
  print(paste("Estimated candidate sampler precision:",round(sdevc,3)))
}

print("Relative bias estimates:")
print(rbc)
print(paste("Mean of relative biases:",round(reffects.mean,3)))
print(paste("Standard deviation of relative biases:",round(reffects.sdev,3)))
print(paste("Estimated probability relative bias exceeds DQO limits (Normal
assumption):",round(prob.rb,3)))
```