ESTIMATION OF BIAS AND VARIABILITY OF CONTINUOUS GAS MONITOR DATA: APPLICATION TO CARBON MONOXIDE MONITOR ACCURACY

Stanley A. Shulman and Jerome P. Smith

U.S. Department of Health and Human Services, U.S. Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Applied Research and Technology, 4676 Columbia Parkway - R5, Cincinnati, Ohio 45226.

REPORT DATE:
July, 2001

REPORT NO.:
EPHB 005-128
DPSE 98-030

MANUSCRIPT PREPARED BY:
Debra A. Lipps

To be submitted to the National Technical Information System

(This page for EPHB use only.)

ESTIMATION OF BIAS AND VARIABILITY OF CONTINUOUS GAS MONITOR DATA: APPLICATION TO CARBON MONOXIDE MONITOR ACCURACY

Stanley A. Shulman and Jerome P. Smith

U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Applied Research and Technology, 4676 Columbia Parkway - R5, Cincinnati, Ohio 45226.

ABSTRACT

A method is presented for the evaluation of the bias, variability, and accuracy of gas monitors. This method is based on using the parameters for the fitted response curves of the monitors. Thereby, variability between calibrations, between dates within each calibration period, and between different units can be evaluated at several different standard concentrations. By combining variability information with bias information, accuracy can be assessed. An example using

carbon monoxide monitor data is provided. Since monitor response can be determined in a chamber simultaneously for several monitors, correlation of errors about the fitted lines is considered. Monitor drift between calibrations is also considered. Procedures for variance components estimation are provided for balanced data. Procedures using Proc Mixed in SAS that provide solutions for both balanced and unbalanced data are available from the authors.

Keywords: Accuracy, Monitors, Variance Component Estimation

INTRODUCTION

Direct-reading monitors are available for a wide range of workplace contaminants.[1] Different versions of these monitors serve as fixed monitors for area monitoring in one location, portable monitors for area monitoring in different locations, and as personal monitors. Data from these monitors provide warnings to workers and management of high concentrations of the measured toxic gases, vapors, or aerosols, and enable calculation of average concentrations of these toxic substances. The development of personal direct-reading monitors with data logging capability has increased the usefulness of monitor measurement.

Since the data from monitors can be used for many applications, it is important to determine data accuracy. Extensive protocols for the evaluation and use of direct-reading gas monitors provide procedures for examining performance characteristics such as repeatability, variation in response with temperature and humidity, response time, and long-term stability of response.[2,3] However, these protocols don't allow calculation of overall accuracy of the data from the monitors. For field monitoring methods samples returned to the laboratory for analysis, the accepted National Institute for Occupational Safety and Health

(NIOSH) accuracy criteria requires that with 95% probability the method give results that are within ±25% of the "true value" 95% of the time.[4]

Statistical methods have been developed to determine if monitoring method data adhere to this criterion. In fact, Kennedy et al.[5] gives an extensive procedure for evaluating the accuracy of methods that require instrumental analysis of samples. Direct-reading monitors are different from such instrumental samples because they are calibrated and then used to directly determine air concentrations in the workplace. Thus, evaluation of monitor accuracy requires determinations by the same monitor at several concentrations and repetitions over time. Since Kennedy[5] considers only single determinations that are independent, these results must be extended to cover monitor accuracy. Nevertheless, Kennedy[5] addresses the kinds of questions that are important in monitor evaluation.

In this paper, the application of accuracy criteria to the measurements made by direct-reading air monitors is examined. The method is applied in an accuracy evaluation of carbon monoxide monitor data.

BASIC PROCEDURES FOR ASSESSING MONITOR ACCURACY

An analytical method for continuous monitors or instrumental analysis is evaluated by estimating its accuracy. Accuracy is estimated as a function of estimated variability and bias. The bias is estimated by the ratio of the monitor response to the target concentration. Variability is estimated as the relative standard deviation (RSD) – the ratio of the standard deviation of monitor determinations to the target concentration. These terms are discussed more fully in Kennedy.[5]

To determine the accuracy of data from the direct-reading gas monitor, the response curve of the monitor over a range of concentrations is determined several times over an extended time period. The monitor is calibrated at fixed intervals during the time period. The initial calibration interval can usually be chosen by using the interval recommended by the manufacturer. The concentration range is determined by the range over which the monitor is to be used. Some functional form can be fitted to the monitor's response to concentration; for instance, a straight line or a quadratic curve.

The bias is estimated as the ratio of the estimated mean monitor response to the target concentration response at each concentration. The concentrations are generated to meet a target concentration, but usually the true concentration will differ somewhat from the target. The target concentration is the theoretically determined estimate of the true concentration produced by the generation system. The target concentration may be calculated from the air flows in the generation system or from some independent method for assessing the concentration. It may be that the target does differ from the true concentration. What matters is that over the course of generating the different target concentrations, the true concentration has the target as a mean, and that the deviations from the target are statistically independent. As long as the concentration generated is checked by an independent method to determine that, on average (over many runs), the targets are approximately attained, the method presented here should work. This is important, since if the generation process has a systematic bias, that bias cannot be separated from the monitor bias, since no comparison to an independent method is provided for in these tests. The generation system should also be maintained so that deviations from the target values are relatively small. Kennedy's tests[5] and confidence intervals proposed for the data are justified

when there are errors in meeting the targets, as long as these errors are independently normally distributed.

An appropriate estimate of the variability of monitor determinations must be obtained, which includes components for among-monitor variability and day-to-day variability. These components are not usually included when accuracy of analytical methods is assessed, but are important for assessing monitor accuracy. From estimates of the variance components, the standard deviation of each component relative to the mean monitor response can be determined. The relative sizes of the components can be important in assessing problems. An example assessing the variability over time for laboratory instrumental methods is given in Vecchia.[6] Accuracy assessment requires estimation of total accuracy. For certain designs, it may be possible to estimate different variance components, knowledge of which may be desirable because it enhances control of the total variability. An example of such a situation is discussed in the next section. For analytical methods, the total variability includes components due to analysis and to sampling. These will not be included here, since gas monitors do both sampling and analysis together.

The combined effects of bias and of variability are assessed via accuracy, which is computed so that the probability equals 95% that a measurement by a given monitor will fall within ±A% of the true value. The accuracy(A) is a function of the bias and variability. If the bias and variability are estimated rather than treated as known, then a second probability, corresponding to the chances of obtaining accuracy A associated with the first probability, may be estimated. If the accuracy of the monitor response is sufficient to meet the accepted criteria, then the calibration schedule and procedure chosen can be followed. If not, then a new calibration schedule and procedure can be tried and the study repeated.

QUESTIONS ANSWERED BY STATISTICAL MODELING

The following questions can be answered by the statistical modeling:

a) Is the total variation acceptable?

b) Is the total bias acceptable?

c) Is the accuracy acceptable?

d) Is there a dominant component of variance (time or monitor)?

e) If the bulk of the variability is over time, the following questions may be of interest:

i) If there is large variability within calibration periods, then perhaps the time between calibrations is too long.

ii If there is large variability among calibration periods, it may be that calibration is not as effective as desired in resetting the response of the monitor.

EXPERIMENTAL DESIGNS

The data from an accuracy assessment of monitors can arise from many different designs depending on how many monitors are available for study and the experimental equipment available for the study. If it is possible to study several different monitors in a chamber, then this study can provide a more extensive evaluation of monitor performance. Considerations that arise in such a design are as follows:

1) There can be position effects in the chamber. If I monitors are to be tested at C different concentrations of analyte, then it makes sense to

vary the position of the monitors in some random fashion from concentration to concentration.

2) Monitors exposed to an analyte at the same time will have correlated errors relative to their fitted response curves. The reason is as follows: Since the actual concentration may differ from the target, it will differ from the target in the same way for all monitors exposed at the same time. But it is the target value that is used for the least squares fit of the response curve, and relative to this target value the deviations from the fitted curves tend to be correlated. This correlation will have to be taken into account in the statistical analysis. Also, this correlation is confounded with the factor describing time within calibration period.

3) Regardless of whether monitor responses are made simultaneously, monitors' responses over a wide concentration range tend to have increasing variability of response with increasing concentration. This situation requires using weighted least squares -- weighting, that is inversely proportional to the target or to the square of the target. An important example of this situation is data with a constant RSD.

4) Regardless of whether monitor responses are made simultaneously, it is possible that the responses of monitors within a calibration period may change in a way that varies systematically over time. For instance, their biases may tend to become increasingly larger over time within period, or their variances may follow a particular pattern. Another possibility is that even without some systematic change, there may be autocorrelation involved in the parameter changes from one time to the next.

It is possible to specify several kinds of designs used for the test data:

a) Randomly chosen units of the same monitor type will be periodically recalibrated at approximately the same time. Then they are all handled in a similar way between calibrations, and all are evaluated over several days within each calibration period by exposing them simultaneously in a chamber.

b) Randomly chosen units of the same monitor type will be periodically recalibrated at approximately the same time. Then these monitors are all handled in a similar way between calibrations, but all are evaluated one at a time, at approximately the same time.

c) Randomly chosen units of the same monitor type will be periodically recalibrated, not necessarily at the same time, and are handled differently between calibrations. For instance, rather than storing them in the laboratory, they may be used in the field. Also, they are tested one at a time.

EXAMPLE USE OF PROCEDURE -- CARBON MONOXIDE MONITORS

An accuracy evaluation was done of personal data logging monitors for carbon monoxide.[7] The monitors were the same type: six Drager Model 190 data logging monitors. The monitors were calibrated at 30-day intervals during the period of the evaluation. Calibration consisted of exposure to zero span gas and to 250 parts per million (ppm) CO. The monitor was forced to have its readout agree at the 0 and 250 ppm concentrations. The assumption is then made that the

response in between these two extreme concentrations is linear. According to the manufacturer, the monitor response should be within 3% of the true value for temperatures 50 to 86°F, or within 2 ppm of that value, whichever is greater.

Design (a) above was used for the evaluation. The monitors were placed in a chamber where they were simultaneously exposed to varying concentrations of CO a number of times over an extended time period. The same set of eight concentrations was used during each test exposure. The concentrations to which the monitors were exposed were produced by a gas generation system that employed computer-controlled mass flow controllers to adjust the flow rates of a standard mixture of CO in N_2 and dilution air to produce varying concentrations of CO. Data were collected from the data logging memories of the monitors using the software provided by the manufacturer. The monitor response at each target was recorded after their readings stabilized.

Twelve days of data collected over three months were used in this study. The number of days varied in each month. For each run, six monitors were used at eight concentrations. A quadratic was used to model the monitor responses. Perhaps a straight line was not adequate because of the wide range of data (see

Figure 1 for a plot of one day's determinations by all six monitors). In Figure 1, all determinations made by the same monitor have been connected. The figure indicates small curvature in the response and also the spread of the monitor responses at each concentration. Because the six monitors were used simultaneously at each concentration, the errors in the monitor determinations are expected to be correlated with each other. Plots of the data indicate no apparent systematic biases among the monitors. Within each calibration period and regardless of concentration, the monitors produced lower readings as the number of days since monitor calibration increased. This is indicated by the plot in Figure 2, of the monitor responses at the 100 ppm concentration for all three months of data. Figure 2 also indicates the amount of variability among the three calibration periods. The main variability appears to be between November and the other two months.

STATISTICAL MODELS FOR MONITOR DATA

Accuracy is a function of bias and variability of the measurements made by the monitors. The aim here is the estimation of accuracy from randomly chosen monitors used at randomly chosen times over several calibration periods, at the

beginning of each of which the monitors are calibrated. If the inaccuracy is sufficiently small, then this calibration approach can be adequate. If the inaccuracy is too high and if this is due to variability over time, then modifications can be made to the calibration procedure. As part of this evaluation, an appropriate polynomial form is fitted to each monitor's responses. Ideally, the polynomial would be a straight line, but as indicated for the carbon monoxide example data, a quadratic can be required. In situations where linear response curve is expected but not realized, different calibration procedures might be tried; for instance, calibration over a smaller range.

A model for the monitor response is proposed in (1) below. In the model, the response of a given monitor m is described for day d within a calibration period p at a concentration θ_j. Note that this is not a calibration curve, since monitors are separately calibrated by procedures given by the monitor manufacturer. For instance, in the previous section, a two-point calibration was described for the carbon monoxide monitor. The model in (1) and (2) is for the monitor readings produced in the evaluation experiment. Suppose that each monitor's response in a set may be described by a quadratic at concentration θ_j for calibration period p at day d within the calibration period. For a quadratic response curve, β (the

response vector) is three-dimensional, with components for the intercept $\beta_{pdm,cep}$, the linear parameter $\beta_{pdm,lin}$, and the quadratic parameter $\beta_{pdm,\,quad}$. $Y_{pdm,j}$ is the response of monitor m for calibration period p on day d within this period at concentration θ_j and is given by the following expression:

$$y_{pdm,j} = \beta_{pdm,cep} + \theta_j\,\beta_{pdm,lin} + (\theta_j)^2(\beta_{pdm,\,quad}) + e_{pdm,j} = \mathbf{x}'\,\beta_{pdm} + e_{pdm,j}, \qquad (1)$$

for $\mathbf{x}' = (1, \theta_j, \theta_j^2)$,

where $e_{pdm,j}$ is the error associated with measurement.

This can be arranged in vector and matrix notation. Suppose that for monitor m, on day d of calibration period p, there are 5 monitor responses made at the concentrations θ_1, θ_2, θ_3., θ_4, θ_5. Then the model (1) for y_{pdm} may be written as follows:

for $1 \le p \le P$, $1 \le d \le D$, $1 \le m \le M$,

$$y_{pdm} = \begin{vmatrix} y_{pdm,1} \\ y_{pdm,2} \\ y_{pdm,3} \\ y_{pdm,4} \\ y_{pdm,5} \end{vmatrix} = \begin{vmatrix} 1 & \theta_1 & \theta_1^2 \\ 1 & \theta_2 & \theta_2^2 \\ 1 & \theta_3 & \theta_3^2 \\ 1 & \theta_4 & \theta_4^2 \\ 1 & \theta_5 & \theta_5^2 \end{vmatrix} \begin{vmatrix} \beta_{pdm,cep} \\ \beta_{pdm,lin} \\ \beta_{pdm,quad} \end{vmatrix} + \begin{vmatrix} e_{pdm,1} \\ e_{pdm,2} \\ e_{pdm,3} \\ e_{pdm,4} \\ e_{pdm,5} \end{vmatrix}$$

$$y_{pdm} = \quad X \quad \beta_{pdm} \quad + \quad e_{pdm}$$

where $\beta_{pdm} = \beta + a_p + b_{d(p)} + c_m + ac_{pm} + abc_{pdm}$. (2)

$$\beta_{pdm} = \begin{vmatrix} \beta_{cep} \\ \beta_{lin} \\ \beta_{quad} \end{vmatrix} + \begin{vmatrix} a_{p,cep} \\ a_{p,lin} \\ a_{p,quad} \end{vmatrix} + \begin{vmatrix} b_{pd,cep} \\ b_{pd,lin} \\ b_{pd,quad} \end{vmatrix} + \begin{vmatrix} c_{m,cep} \\ c_{m,lin} \\ c_{m,quad} \end{vmatrix} + \begin{vmatrix} ac_{pm,cep} \\ ac_{pm,cep} \\ ac_{pm,quad} \end{vmatrix} + \begin{vmatrix} abc_{pdm,cep} \\ abc_{pdm,lin} \\ abc_{pdm,quad} \end{vmatrix}$$

for the quadratic model. Aside from β each of the addends in (2) is a vector that represents variability of the components of β_{pdm} over periods (p), days (d), or monitors (m) or combinations of these. If the parameters of interest were means, rather than the parameters associated with a curve, each of the addends in (2) would be one dimensional. For instance, suppose that m monitors were used to measure a property of a product, on day d within period m. If the interest were in

the variability of the product over monitors, between periods, and over days within period, then model (2) might be appropriate, except that each component would be one dimensional. This one-dimensional form is a form that often arises in quality control.

In the above version, the subscript "cep" denotes components of the intercept of the response curve, "lin" denotes components of the linear parameter of the response curve, and "quad" denotes components of the quadratic parameter of the response curve. If there were T combinations of calibration periods and days within these periods, and M monitors, then there would be TM models of the form (1) for the quadratic response curve model.

The total variance of a measurement, y_{pdm}, may be seen to be the sum of its components:

$$\Sigma_T = \mathrm{Var}(y_{pdm}) = (X)V_\beta(X)' + \mathrm{Var}(e_{pdm}), \text{ where} \qquad (3)$$

$$V_\beta = V_P + V_{PD} + V_M + V_{PM} + V_{PDM}, \text{ and } \mathrm{Var}(e_{pdm}) = (\sigma_e)^2 D, \text{ a diagonal matrix,}$$

where the variance associated with each component is subscripted by the upper case letter corresponding to that component. Also, X' denotes the transpose of the

matrix X. Note that V_β does not include Var(e_{pdm}), since the first is the variance matrix associated with the randomly chosen β_{pdm}s, and the second is the variance associated with the deviation of the actual measurements from this randomly chosen curve. For instance, V_p denotes the variance matrix of average monitor determinations over calibration periods. For the one-dimensional quality control example mentioned above, each addend on the right side of (3) would be one-dimensional, a variance associated with that component. When a quadratic is fitted, since three parameters are varying, each addend is a three-by-three matrix. On the diagonals are the parameter variances associated with that component (whether p,d,m or combinations of these), and the off-diagonals are for the covariances. The representation given in (3) only makes sense if all components of the model (2) above are statistically independent. Otherwise, there will be terms in (3) for the covariance between different components.

Some important statistical assumptions pertaining to the above models are now stated:

Expectation (β_{pdm}) = β, Variance (β_{pdm}) = V_β,

Expectation (e_{pdm}) = 0, Variance (e_{pdm}) = $(\sigma_e)^2$ D, for a diagonal matrix D,

Covariance (β_{pdm}, e_{pdm}) = 0.

β is the only fixed factor in the equations of (2), and other components are taken as normally distributed, and statistically independent. β is the vector of mean parameters for the monitor response curve, averaging over monitor (m), calibration period (p), and day within calibration period (d). β_{pdm} is the parameter vector of the monitor response curve for monitor m on day d of calibration period p. Suppose that measurements are made at n different concentrations for monitor m on day d of calibration period p. In the notation used above, y_{pdm} is the n x 1 dimensional vector of monitor responses of monitor m on day d of calibration period p. X is an n x q matrix, where (q-1) denotes the degree of the polynomial which describes the response function. Because the same design matrix is used at all evaluation times, it is assumed that the same concentrations are used for each evaluation of the monitors. Although the statistical approach used here could be presented without this restriction, this simplification makes the notation a little easier to follow. Also, if the same concentrations were not used, and the range from lowest to highest concentration varied considerably from one evaluation to another, this could require a more complicated model than is presented here. There is more discussion of this in the appendix.

As in the univariate version of the random effects model, the random components are assumed to be normally distributed, and each random component is assumed to be statistically independent of every other component. See Scheffe,[8] in which the independence of the times and monitors, and the normality assumptions on the errors are used to imply the independence of all error terms. For any of the designs mentioned above, there may be situations in which the day within calibration period component cannot be viewed as a random sample. This is true of the example data, as is clear from Figure 2, which suggests trends in monitor response during each calibration period. The sample program discussed in Appendix II and available from the authors allows for adjustment for trends within calibration period.

The component e_{pdm} includes measurement error associated with the monitors and with the test concentrations used in the evaluation. In many situations, it makes sense to assume that the measurement error is statistically independent of the components of β_{pdm}. The rationale is that the changes in the components of the response curve are gradual, but the measurement errors are instantaneous, and each one represents the random noise associated with the monitor determination. This assumption of independence is also a reason that the component e_{pdm} is

written separately in the model (1). This assumption of independence is a common assumption in random effects models,[6,8,9] but there could be situations in which it is unreasonable. Notice that measurement error pertains only to the y_{pdm} and not to the β_{pdm}, since the ys are observations and the βs are not.

In many situations, it may make sense to assume that the

$$\text{Variance } (e_{tm,j}) = (\sigma_e)^2 (\theta_{pdm,j})^2.$$

$$\text{Thus, } D = \text{Diagonal } [(\theta_{pdm,1})^2 \ldots (\theta_{pdm,J})^2].$$

ESTIMATION OF ACCURACY

The estimates of the mean and variances for the above model will be based on weighted least square estimates obtained for each response curve by each monitor, each time that the monitor is evaluated. For accuracy estimation, the important quantities are the variance of y_{pdm}, the estimated average bias relative to the target value, and the variance of the average bias. Let $b_{pdm,LS}$ be the estimated parameter vector (including intercept, linear parameter, and quadratic parameter) obtained

via weighted least squares for data obtained from monitor m on day d of calibration period p. Then

$$b_{pdm,LS} = (X'D^{-1}X)^{-1}X'D^{-1}y_{pdm}, \quad (4)$$

where the weighting diagonal matrix D accounts for the dependence of the variance of the y_{pdm} on the concentration values $\Theta_{pdm,k}$. For instance, if the variance is proportional to Θ_k^2, the kth diagonal is Θ_k^2. X is the design matrix for the response curve, assumed to be based on the same concentrations each time a response curve is fitted. The model for b_{pdm} can be rewritten in terms of $b_{pdm,LS}$:

$$b_{pdm,LS} = \beta + a_p + b_{pd} + c_m + ac_{pm} + abc_{pdm} + f_{pdm}, \quad (5)$$

where $f_{pdm} = (X'D^{-1}X)^{-1}X'D^{-1}e_{pdm}$,

where the component f takes into account the variability of the data about the fitted least squares curve, and is a function of the measurement error. Because there is just one $b_{pdm,LS}$ for each calibration period, day in period, and monitor, the two components abc_{pdm} and f_{pdm} cannot be separately estimated. However, the

residuals from the least square estimators $b_{pdm,LS}$ can be used to obtain a pooled estimate $(\hat{\sigma}_e)^2$ of $(\sigma_e)^2$. Because of (2), (3), and (5),

$$V_{\beta,LS} = Var(b_{pdm,LS}) = V_p + V_{d(p)} + V_m + V_{pm} + V_{pdm} + \sigma_e^2(X'D^{-1}X)^{-1}, \qquad (6)$$

and $Var(\beta_{pdm}) = V_\beta = V_{\beta,LS} - \sigma_e^2(X'DX)^{-1}$, with estimators $\hat{V}_\beta$ for V_β, and $\hat{V}_{\beta,LS}$ for $V_{\beta,LS}$, and $(\hat{\sigma}_e)^2$ of $(\sigma_e)^2$. The estimators $\hat{V}_{\beta,LS}$ and $(\hat{\sigma}_e)^2$ are statistically independent, which is useful when the degrees of freedom associated with the estimate of the variance of $y_{pdm,k}$ must be estimated.

The estimates of average bias and total relative standard deviation of a single measurement can then be obtained.

We first discuss this for the situation where the data are balanced, by which we mean that all monitors are used at each evaluation, and there are equal numbers of evaluation days in each calibration period. Let B_{avg} denote the average of all the least square parameter estimates. The rationale for this estimator is described in Reference 9.[9] B_{avg} is the average of all the estimates $b_{pdm,LS}$. Then the Bias B* at target concentration Θ is estimated by $\hat{B}^*$.

$$\hat{B}^* = x'^* B_{avg} / \Theta, \text{ where } x' = (1 \ \Theta \ \Theta^2). \quad (7)$$

Kennedy[5] refers to B* as the unsigned bias, whereas B = B*-1 is called the bias.

The variance of the determination at Θ, the $y_{pdm,j}$ of (1), is estimated by $x'^*(\hat{V}_{\beta,LS})^*x + (\hat{\sigma}_e)^2 [D(k,k) - x'(X'DX)^{-1}x]$, where D(k,k) is the kth diagonal element of the weight matrix D (see (A7) of Appendix I). Then the relative standard deviation, S_{rT}, of a determination at Θ is estimated by

$$\hat{S}_{rT} = \{x'^*(\hat{V}_{\beta,LS})^*x + (\hat{\sigma}_e)^2 [D(k,k) - x'(X'DX)^{-1}x]\}^{.5} / x'B_{avg}. \quad (8)$$

(8) gives the estimated relative standard deviation of an actual instrumental determination. For accuracy, it is easiest to place a confidence band on the standard deviation relative to the target, $\hat{S}_{rT,tr} = \hat{S}_{rT} \hat{B}^*$. More detail is given about these procedures in Appendix I. The estimator $\hat{V}_\beta$ depends on the experimental design. That shown in (A7a) of Section 1 of Appendix I is appropriate for the design used in the example. The degrees of freedom associated with (8) will be denoted by r2. For balanced data a spreadsheet is available from the authors. (See Appendix II).

Often the data will not be balanced. Appendix II discusses the SAS[10] procedures to obtain estimates even when the data are not balanced. If they are balanced, then the approach will yield estimates close to the estimates shown in the Appendix I. The SAS program is available from the authors.

If the data are unbalanced, then the estimators given in (A7) of Appendix I, Section 1, cannot be used. A solution can be obtained by working with the $\hat{y}_{pdm,j} = x' b_{pdm,LS}$ (x' given in (1)), which correspond to the estimated monitor responses at θ_j. The variance components for these $\hat{y}_{pdm}$s can be obtained by using a procedure in SAS called Proc Mixed[10]. This procedure can be used whether the data are balanced or unbalanced. (See Section 1 of Appendix I for more explanation.)

Whether the data are balanced or unbalanced, the estimates of average bias and total relative standard deviation can then be used to obtain accuracy estimates. It is useful to have an approximation of the accuracy function, since the exact solution of the integral equation defining accuracy cannot be written out explicitly. An approximation developed by R. Song[(11)] can be used. The approximation is intended to give the accuracy A such that the chances are 95%

that a measurement will be between $\theta(1-A)$ and $\theta(1+A)$, for true value θ, assuming known bias and relative standard deviation. For estimated bias and estimated relative standard deviation, (9) provides an estimate of 95% accuracy.

$$A(\hat{B}^*,\hat{S}_{rT,tr}) = (1.96 - 0.39)[\hat{S}_{rT,tr}] + [(0.39\ \hat{S}_{rT,tr})^2 + (\hat{B}^*-1)^2]^{0.5} \quad (9)$$

SUMMARY OF ASSUMPTIONS

If certain assumptions about the data are met, then standard computer programs can be used to compute the variance components. The main assumptions are as follows:

1) A straight line or quadratic describes the monitor response, although weighted least squares may be needed to attain constant variance of the residual error.

2) The weighted residuals should be approximately normally distributed.

3) If design a) specified above has been used, then any correlation among measurements must be small compared to the other sources of error.

4) Plots should be used to assess the possibility of systematic changes of monitor response within calibration period.

When the above assumptions are met, the SAS code discussed in Appendix II and available from the authors can be used.

DISCUSSION OF ADVANTAGES AND DISADVANTAGES OF PROPOSED METHOD

The main statistical problem in this work is the estimation of the variances required for accuracy evaluation. For balanced data (equal number of days in each period), the statistical approach used here produces estimates by a method similar to that used for one dependent variable, rather than the two- or three-dimensional variable required to describe the response curve. This is explained more fully in Appendix I. The estimation method used is a methods of moments method, in which the mean squares from the analysis of variance are equated to their

expected values. After the solutions are obtained for the variances, it is the predicted values from the model that are used in accuracy, by producing estimates of accuracy at one concentration at a time.

The method shown in Section 1 of Appendix I is for balanced data. For such data a spreadsheet is available from the authors. (See Appendix II.) The approach may also be implemented by using Proc Mixed in SAS, in which the restricted maximum likelihood estimation procedure gives the same solution for balanced data as the approach in Appendix I gives. However, Proc Mixed provides estimates whether or not the data are balanced, and also permits assessment of trends within calibration periods.

Since Proc Mixed provides solutions whether or not the data are balanced, why present the method of moments estimates in Appendix I? There are two reasons. First, the method of moments estimates are closed form, and are useful for assessing the effect of correlation among the errors for simultaneous evaluation of monitors. This correlation could arise because of deviations from the target concentrations to which monitors are exposed. Results pertaining to such errors have been worked out by the authors, but are not presented here. A second reason

for presenting the matrix formulation in Appendix I is that this formulation can probably be used to place a confidence interval on the monitor accuracy for all concentrations in an interval, instead of only a few chosen concentrations. This extension of the results has not been developed for this paper.

EXTENSION OF PROCEDURE TO ACCOUNT FOR EXPLANATORY VARIABLES

The procedure can be extended to examine the effects of other variables such as humidity, temperature, and interferences on the accuracy of the measurements made by the monitors. This is accomplished by determining the response function of the monitors over a wide concentration range of temperature and humidity. This should be done a number of times over a long time period, just as in the example that is given in this paper. The model can be extended to examine these effects by including terms in the model (2) that are functions of these additional variables. The SAS computer code can be modified to include these new variables.

ANALYSIS OF EXAMPLE DATA

Since each of the 12 days of data described earlier in the section "Example Use of Procedure" contributed six sets of parameter estimates (one for each of the six monitors), there were 72 fits of the quadratic response curve. For the three months of the example data, the largest component of variability is among months. Since this is the largest component, the degrees of freedom of the estimated total variance is closest to the degrees of freedom of this largest component -- between 3.00 and 3.94 for concentrations between 10 and 50 ppm. For these data, the estimated average parameter values and total variance matrix are as follows:

B_{avg}= (-1.785 1.052 -0.000379)', where -1.785 is the average value for the intercept estimate, 1.052 is the average value for the linear parameter estimate, and -0.000379 is the average value for the quadratic parameter estimate, estimates that were obtained by weighted least squares applied to each set of eight measurements for each monitor. Results change little if a narrower range of data is used, say, the lower six test concentrations. In the notation of equation (3), an estimate of $V_{\beta,LS}$ is:

$$\begin{vmatrix} +1.29 & -0.0144 & +0.0000581 \\ -0.0144 & +0.00850 & -0.0000208 \\ +0.0000581 & -0.0000208 & +7.33x10^{-8} \end{vmatrix}, \qquad (10)$$

where the diagonal terms are the variance estimates of the parameter estimates, and the off-diagonal terms are the covariances. The matrix $\hat{V}_{\beta,LS}$ is used in computing the total relative standard deviation given in (8), required for accuracy estimation. The estimates are derived from the full model (2), from which no variance components have been removed.

Assumption (3) in the section "Summary of Assumptions," requires that any correlation among measurements must be small compared to the other sources of error. For the example carbon monoxide monitors, the estimated ratio of $(\sigma_\delta)^2$ to $(\sigma_e)^2$ is about 0.1715. $(\sigma_\delta)^2$ must be large to warrant modification of the simple formulas (A7 of Section 1, Appendix I) for the expected values of the mean squares. Since$(\sigma_e)^2$ is relatively small compared to the other variances, $(\sigma_\delta)^2$ is much smaller. Thus, we believe that it can be ignored.

Since much of the interest is at low concentrations, we give the accuracy estimates and individual 95% bands, as discussed in Section 4 of Appendix I, for concentrations of 10, 30, and 50 ppm. These estimates are shown in Table I.

Notice that in Table 1 the confidence limit for the bias is the 2.5% limit, and that for $S_{rT,tr}$ is the 97.5% limit. When substituted into the accuracy equation (10), an upper 95% confidence limit for accuracy is obtained.[5] The accuracy figures in the second column of the table are similar (although somewhat larger) to the manufacturer's specifications, while accuracy limit figures in the last column above indicate that the accuracy based on the evaluation experiment is considerably poorer than that provided by the manufacturer - that the monitor response should be within 3% of the true value for temperatures 50 to 86°F, or within 2 ppm of that value, whichever is greater. However, it should be pointed out that the manufacturer's specifications are most likely based on single measurements made on the same day that the monitors are calibrated, while the estimates here are measurements made over a considerable time period where the monitors are calibrated at the beginning of the time period. Monitors tend to underestimate the concentration, though the size of the underestimate varies with the concentration. The upper limits on accuracy may be much bigger than the

accuracy estimate because the main sources of variability, calibration period (accounting for over half the variance of intercept and linear parameter), had few degrees of freedom -- 2, or because the true accuracies are higher than the point estimates. More months of data and more monitors might have lowered the upper limits on accuracy for this experiment, unless new data indicated higher point estimates.

As was pointed out in the previous discussion of Figure 2, there appears to be a loss of sensitivity of these monitors as days since calibration increases. The model used here, which is discussed more fully in Appendix I, assumes that the days within calibration period are treated as a random sample, as far as the variability of the response curve parameters is concerned. This may not be an appropriate assumption in this situation.

The statistical analysis does indicate that at the 30 and 50 ppm concentrations, there is statistically significant decrease with day since calibration. When this is allowed for, the upper confidence limits for accuracy improve somewhat for analyses carried out at about 10 days since calibration (the average of the days since calibration for which there are data). For 10, 30, and 50 ppm, the new upper

accuracy limits are 1.23, 0.594, and 0.456, which are improvements over the 1.25, 0.669, and 0.549 shown in the table above. Improvement is to be expected, since by allowing for dependence on day since calibration, we reduce the variances used in accuracy calculation.

In summary, bias tends to be negative, as high as -13% at 10 ppm, and relative standard deviation as high as 14% at 10 ppm. For 10 to 50 ppm, accuracy estimates >2 ppm limit provided by manufacturer, and exceed 12 ppm when variability of estimate is taken into account. Accuracy limits are wide, because more calibration periods should have been studied.

CONCLUSIONS

A method is presented for the evaluation of the bias, variability, and accuracy of gas monitors. This method is based on using the parameters for the fitted response curves of the monitors. Thereby, variability between calibrations, between dates within each calibration period, and between different units can be evaluated at several different standard concentrations. Although the mathematics involved is somewhat complicated, for the case of balanced data, calculations can

be done on a spreadsheet. This should enhance the applicability of the method. The spreadsheet is available from the authors. The appeal of the method is that it allows evaluation of monitor precision, bias, and accuracy at many different concentrations, and can be generalized to evaluate the effect of additional variables, such as temperature and humidity, in the evaluation. SAS computer code to implement the approach for unbalanced data is available from the authors. An example has been provided for assessment of accuracy of carbon monoxide monitors.

ACKNOWLEDGMENTS

We wish to thank reviewers for their helpful comments: Thomas Fischbach, Martin Petersen, Paul Hewett, Anne Votaw, Dennis O'Brien, David Bartley, and Martin Abell of NIOSH for their reviews of the paper, and Teresa Seitz of NIOSH for her review of a poster session[12] based on this material. Thanks to Debbie Lipps for preparing the manuscript.

APPENDIX I

1. ESTIMATION OF VARIANCE COMPONENTS BY ANALYSIS OF VARIANCE METHOD - BASIC FORMULAS WHEN ZERO TARGET ERRORS ARE ASSUMED AND DATA ARE BALANCED

It seems easiest to present a simplification of the model (1) for the parameter vector β_{pdm}, say β_{tm}, where there is just one time variable indexed by t:

$$\beta_{tm} = \beta + a_t + c_m + ac_{tm}, \tag{A1}$$

$$y_{tm} = X\,\beta_{tm} + e_{tm},$$

$$Var(\beta_{tm}) = V_t + V_m + V_{tm},$$

$$Var(y_{tm}) = X' Var(\beta_{tm}) X + Var(e_{tm}),$$

$$\text{Expectation}\,(\beta_{tm}) = \beta,\ \text{Variance}(\beta_{pdm}) = V_\beta,$$

Expectation (e_{tm}) = 0, Variance(e_{tm}) = $(\sigma_e)^2$ D, for a diagonal matrix D, Covariance (β_{tm} ,e_{tm}) = 0, Covariance (a_t,e_{tm}) = 0, Covariance (c_m,e_{tm}) = 0, Covariance (ac_{tm}, e_{tm}) = 0, and all covariances among the components of β_{tm} are 0.

Since all components are assumed to be normally distributed, the zero covariance assumptions above imply that each component is statistically independent of every other component. In this section, we act as if there are no errors in meeting the targets. In many situations, it may make sense to assume that the variance ($e_{tm,j}$) = $(\sigma_e)^2$ $(\theta_{tm,j})^2$. Thus, D is a JxJ diagonal matrix:

$$D = \text{Diagonal } [(\theta_{tm,1})^2 \ldots (\theta_{tm,J})^2]. \qquad \text{(A2)}$$

Let $b_{tm,LS}$ be the single parameter vector for the least square estimate of monitor m used at time t. Assume a common design matrix X for all combinations of t and m. Then $b_{tm,LS} = (X'D^{-1}X)^{-1}X'D^{-1}y_{tm}$, where the weighting diagonal matrix D is needed to account for the dependence of the variance of the y_{tm} on the concentration values x_{tm}, and the model (A1) can be rewritten:

$$b_{tm,LS} = \beta + a_t + c_m + ac_{tm} + f_{tm}, \qquad \text{(A3)}$$

where $f_{tm} = (X' D^{-1} X_T)^{-1} X' D^{-1} e_{tm}$, where e_{tm} is distributed $N(0,\sigma_e^2 D)$, and where it is necessary to include this new component f to take into account the variability of the data about the fitted least square curve. This new component is independent of the other components since e_{tm} is.

For estimation of variance components, the following sums of squares are used.[8]

$$(M-1)*MS_m = T \Sigma_m (b_{m,LS} - B_{avg})*(b_{m,LS} - B_{avg})' \qquad \text{(A4)}$$

$$(T-1)*MS_t = I \Sigma_t (b_{t,LS} - B_{avg})*(b_{t,LS} - B_{avg})'$$

$$(MT-1)*MS_{Tot} = \Sigma_{tm} (b_{tm,LS} - B_{avg})*(b_{tm,LS} - B_{avg})',$$

$$(M-1)(T-1)*MS_{tm} = (MT-1)*MS_{Tot} - (M-1)*MS_m - (T-1)*MS_t,$$

where M is the number of monitors used, T is the number of times that each monitor is used. A balanced design is used here, where it is assumed that all monitors are used at each time. B_{avg} denotes the average of the least square parameter estimates, where the average is taken over all monitors and times.[9]

B_{avg} is normally distributed with mean β and variance

$V_t/T + V_m/M + [V_{tm} + \sigma_e^2 (X_T' D^{-1} X_T)^{-1}]/(TM)$.

The statistical expected values can be obtained by noting that the same multipliers that pertain to each component of the mean square matrices and the variance matrices V apply to the entire matrix. The one difference is that $\beta_{tm,LS}$ differs from β_{tm} by the statistically independent addend f_{tm}, due to least squares. Because of statistical independence and because $\beta_{tm,LS} = \beta_{tm} + f_{tm}$, the expected values of mean squares must have the additional addend $\sigma_e^2(X' D^{-1} X)^{-1}$, where it is assumed that the same concentrations are used in each evaluation. If not, then the average value of $\sigma_e^2(X' D^{-1} X)^{-1}$ over all evaluations is used in place of $\sigma_e^2(X' D^{-1} X)^{-1}$.

$$E(TM-1)MS_{Tot} = (TM-1)V_{\beta,LS} - (M-1)V_t - (T-1)V_m \qquad (A5)$$

$$EMS_m = TV_m + V_{tm} + \sigma_e^2 (X' D^{-1} X)^{-1}$$

$$EMS_t = MV_t + V_{tm} + \sigma_e^2 (X' D^{-1} X)^{-1}$$

$$EMS_{tm} = V_{tm} + \sigma_e^2 (X' D^{-1} X)^{-1}$$

The above expected values are based on assumptions of 0 errors in targets, which means no correlation in errors of monitors evaluated simultaneously. Also, the data are assumed to be balanced.

In the above expressions for the expected values of the mean squares, the addend $\sigma_e^2 (X' D^{-1} X)^{-1}$ must be replaced by $(1/T)\Sigma_t \sigma_{et}^2 (X_t' D^{-1} X_t)^{-1}$ if the design matrix or the residual variance about the fitted curve varies with the time t. If the evaluations have very different ranges between lowest and highest concentration, then this could also cause inhomogeneity in the components of $V_{\beta, LS}$. If the inhomogeneity is ignored, the estimates of the variances of the fitted values of the instrumental response curve will be biased. Since the concentrations are set by the experimenter, it is sensible to try to use the same test concentrations at each evaluation. If there were some evaluations where not all concentrations were evaluated, the experimenter should see that this does not unduly affect the estimates of β and the variance $\sigma_e^2 (X' D^{-1} X)^{-1}$of β.

In the analysis of variance method for estimating variance components, mean squares, computed from the data, are equated to their expected values. Estimates are then obtained by solving these equations. Let $\hat{V}_C$ denote the estimate of V_C for component C. Then

$$\hat{V}_{tm} = MS_{tm} - Est.\sigma_e^2(X' D^{-1} X)^{-1} \quad (A6)$$

$$\hat{V}_t = (MS_t - MS_{tm})/M$$

$$\hat{V}_m = (MS_m - MS_{tm})/T$$

$$\hat{V}_{\beta,LS} = \hat{V}_t + \hat{V}_m + MS_{tm}$$

The above estimators can next be used to obtain estimates for the quantities needed for accuracy estimation. In (A7) $y_{tm,j} = x`\beta_{tm} + e_{tm,j}$, for $x` = (1, \theta_j, \theta_j^2)$.

a) $$\hat{V}_\beta = \hat{V}_{\beta,LS} - Est.\sigma_e^2 (X' D^{-1} X)^{-1}, \quad (A7)$$

b) $$Est.\ Var.\ (y_{tm,j}) = x'\hat{V}_\beta x + Est.\sigma_e^2\, x' (X' D^{-1} X)^{-1} x\ ;$$

c) $$\hat{S}_{rT,tr} = [Est.\ Var.(y_{tm,j})]^{0.5}/x'\, B_{avg}$$

d) $$\hat{B}^* = \text{Estimated Unsigned Bias} = x'\, B_{avg}/\Theta,$$

$$\text{for } B_{avg} = \sum_{t=1}^{T} \sum_{m=1}^{M} b_{tm,LS}/(TM),$$

e) $$\hat{S}_{rT,tr} = [Est.\ Var.\ (y_{tm})]^{0.5}/(x'\, B_{avg})$$

f) $$\hat{\sigma}^2(\hat{B}^*) = Est.\ Var.\ (Est.\ Bias)$$

$$= (1/\Theta^2)\, x'\{\hat{V}_t/T + \hat{V}_m/M + [\hat{V}_{tm} + \hat{\sigma}_e^2(X' D^{-1} X)^{-1}]/(TM)\}x,$$

for $x' = (1\ \Theta_j\ \Theta_j^2)$ and Est.σ_e^2 an estimate of σ_e^2. The formula b) gives an estimate of the variance of a future determination by a randomly chosen monitor. The above estimators are appropriate for designs a) and b) discussed in "Experimental Designs" above.

When the data are analyzed in Proc Mixed, the dependent variable is $\hat{y}_{tm,j} = x'\, b_{tm,LS} = x'\beta + x'\, a_t + x'c_m + [x'ac_{tm} + x'f_{tm}]$, from (A3). Components with different subscripts are statistically independent. From (A4),

$$x'[(M-1)*MS_m]x = T\,\Sigma_m\, x'\,(b_{m,LS} - B_{avg})*(b_{m,LS} - B_{avg})'\,x$$

$$= T\,\Sigma_m\ (x'\,b_{m,LS} - x'\,B_{avg})*(x'b_{m,LS} - x'\,B_{avg})'$$

$$= T\,\Sigma_m\ (\ \hat{y}_{m,j} - x'\,B_{avg})*(\ \hat{y}_{m,j}\ - x'\,B_{avg})',$$

for $\hat{y}_{m,j} = \Sigma_t\ \hat{y}_{tm,j}/T$. Analogous relations apply to the other mean squares in (A4). These are the mean squares of the one dimensional model based on the $\hat{y}_{tm,j}$s. Thus, the estimators given in (A6) can be converted to those for the one-dimensional model by premultiplying by x' and postmultiplying by x. For balanced data the means squares $x'\,M_{sc}\,x$ are statistically independent, and are independent of the average value $\Sigma_{tm}\,\hat{y}_{tm,j}/(TM)$.

2. IMPORTANCE OF ZERO TARGET ERROR ASSUMPTION

Errors in meeting the targets are important because they lead to bias in the expectations of the mean squares. When there are no errors or negligible errors, it seems reasonable to treat the correlation of the y_{tm}s made at the same time t as negligible. If the generated values are quite close to target values, then this may be reasonable. From the data itself, we may reach a decision.

The presence of non-negligible error causes a new component g_{tm} to appear in the model:

$$b_{tm,LS} = \beta + a_t + c_m + f_{tm} + g_{tm}.$$

g_{tm} has terms that express the deviation of the realized concentration values from the target values. These deviations are multipliers for the parameters in β,a_t, and c_m. As is shown in Box,[13] for the linear response curve, the expected value of g_{tm} is 0, but for quadratic response curves it is not 0. The new term causes expected values of mean squares to differ from those given in (A6). For balanced data, calculations can be done to obtain correct expected values of the mean squares, if

errors in target and the errors around the fitted least square curve can be assumed statistically independent.

3. ESTIMATING THE DEGREES OF FREEDOM

The estimation of the total variance also requires the estimation of the degrees of freedom of the estimate. For the two-way layout, as long as there are no missing observations, the data can be viewed as balanced, which means that all TM cells in the two-way table have the same number of observations per cell. More generally, balanced data require that for any combination of levels of the explanatory factors (for instance, time and monitor), the same number of repeats are present for each such combination of factor levels.

When the data are balanced, the mean squares of (A5) evaluated at $x' = (1\ \theta_j\ \theta_j^2)$ are statistically independent. Also the pooled estimator of σ_e^2 can be shown to be statistically independent of the other mean squares. When the mean squares are statistically independent and follow chi-square distributions, Satterthwaite's approximation can be used to estimate the degrees of freedom.[14] For a linear combination of independently distributed mean squares, M_i, arising from normally

distributed data, with d_i degrees of freedom, i = 1 to m, the estimated degrees of freedom of the estimated variance of $y_{tm,j}$ (monitor m determination at time t at θ_j) is $s^2 = \Sigma_i k_i M_i$:

$$r = s^2/[\Sigma_i (k_i M_i)^2/d_i], \qquad \text{(A8)}$$

and rs^2/(true variance of $y_{tm,j}$) is approximately distributed as $\chi^2(r)$. This formula for r can be used to calculate Var. ($y_{tm,j}$), as given in (A7)(b).

Even when the total number of items sampled is large, the estimate of r can be small if the largest contribution is from a component that has few degrees of freedom.

For unbalanced data, Proc Mixed makes use of a generalized Satterthwaite approximation.[10,15]

4. CONFIDENCE LIMITS FOR MONITOR ACCURACY

It is much easier to obtain an estimate for accuracy than to obtain confidence limits for the true accuracy. As is mentioned in the main part of the text, estimated accuracy can be obtained by substituting estimates of bias and relative standard deviation into the equation for accuracy or into the approximation developed by Song.[10] The method used here for placing confidence limits on accuracy is the following, based on Bonferroni inequalities:[16]

Obtain confidence limits for bias and standard deviation, under normality assumptions (given after equation (3) above). The confidence limit for the bias is the 2.5% limit if $(x_j' B_{avg}) < \theta_j$ or the 97.5% limit otherwise [5], and that for $S_{rT,tr}$ is the 97.5% limit, used in (A9). When substituted into the accuracy equation (A10), an upper (no less than) 95% confidence limit for accuracy is obtained. This method is conservative, especially so for small degrees of freedom of the relative standard deviation estimate.

$$x_j' \beta/\theta_j > [x_j' B_{avg} + t_{r1}^{.025}[\text{Est. Var. } (x_j' B_{avg})]^{.5}/\theta_j \text{ if } (x_j' B_{avg}) < \theta_j,$$

$$x_j' \beta/\theta_j < [x_j' B_{avg} + t_{r1}^{.975}[\text{Est. Var. } (x_j' B_{avg})]^{.5}/\theta_j \text{ if } (x_j' B_{avg}) > \theta_j,$$

$$(S_{rT,tr})^2 = [Var(y_{tm})/\theta_j^{\,2}] < (r2)(\text{Est. Var. } (y_{tm})/\theta_j^2)/(\chi^{\,2}(.025,r2), \qquad \text{(A9)}$$

for $x_j' = (1\ \Theta\ \Theta^2)$, and $(\hat{S}_{rT,tr})^2$ is an estimate of the total variance of a determination y_{ti} relative to Θ. Thus, $\hat{S}_{rT,tr} = \hat{S}_{rT}\hat{B}^*$. Satterthwaite's approximation can be used to calculate r1 and r2. R2 and r1 are, respectively, the degrees of freedom associated with $\hat{S}_{rT,tr}$ and with the estimated average bias. Substitution of the confidence limits from (A9) into (A10) in place of $\hat{B}^*$ and $\hat{S}_{rT,tr}$ gives the upper confidence limits for accuracy.[5]

$$A(\hat{B}^*,\hat{S}_{rT,tr}) = (1.96 - 0.39)[\hat{S}_{rT,tr}] + [(0.39\ \hat{S}_{rT,tr})^2 + (\hat{B}^*-1)^2\ ]^{0.5} \qquad \text{(A10)}$$

The computer code discussed in Appendix II computes the above accuracy confidence limit.

APPENDIX II

SAS CODE FOR EVALUATING MONITOR ACCURACY

For balanced data a spreadsheet is available from the authors. For balanced or unbalanced data, a SAS program has been written that fits a quadratic curve for each monitor evaluation. The program assumes that all fitted curves are fitted at the same concentration values. This program is appropriate for either designs a or b discussed in the section above, "Experimental Designs." The required parameter estimates are obtained in two steps. The program first uses PROC REG to fit the quadratic response curve on day d of period p for monitor m. The resulting parameter estimates are used by PROC MIXED to obtain the variance components and average parameter estimates needed for accuracy analysis. Similar results could also have been obtained in one step by fitting all the monitor measurements to a quadratic response curve in PROC MIXED, allowing for appropriate estimates of random effects. A reason to prefer the two-step procedure is that there may sometimes be convergence problems in the larger data set associated with the one-step procedure. In other words, PROC MIXED may not give a solution, and the model may have to be modified so that a solution can be obtained from a simpler model. The two-step procedure is also consistent with

the procedure shown in Section 1 of Appendix I for obtaining estimated variances for balanced data.

REFERENCES

1. **Cohen, B., and S. Hering, technical eds.**: *Air Sampling Instruments for Evaluation of Atmospheric Contaminants*. 8th ed. Cincinnati, Oh.: ACGIH, 1995.

2. **Instrument Society of America (ISA)**: *Performance Requirements for Carbon Monoxide Detection Instruments (50-1000 ppm Full Scale)*, Document No. ISA-S92.02.01-1998, Part I. Research Triangle Park, N.C.: ISA, 1998.

3. **Instrument Society of America (ISA)**: *Installation, Operation, and Maintenance of Carbon Monoxide Detection Instruments (50-1000 ppm Full Scale)*, Document No. ISA-S92.02.02-1998, Part II. Research Triangle Park, N.C.: ISA, 1998.

4. **Busch, K.A., and D.G. Taylor**: Statistical Protocol for the NIOSH Validation Tests. In *Chemical Hazards in the Workplace -*

Measurement and Control. G. Choudhary, ed. Washington, D.C.: American Chemical Society Symposium Series 149, 1981.

5. **National Institute for Occupational Safety and Health**: *Guidelines for Air Sampling and Analytical Method Development and Evaluation*, by E. Kennedy, T. Fischbach, R. Song, P. Eller, and S. Shulman (DHHS/NIOSH Pub No. 95-117). Washington, D.C.: Government Printing Office, 1995. pp. 1-4, 44-48.

6. **Vecchia, D.F., H. Iyer, and P.L. Chapman**: Calibration with Randomly Changing Standard Curves. *Technometrics* 31(1):83-90 (1989).

7. **Smith, J.P., and S. Shulman**: Evaluation of a Personal Data Logging Monitor for Carbon Monoxide. *Appl. Occup. Environ. Hyg.* 9(6):418-427 (1994).

8. **Scheffe, H.**: *The Analysis of Variance*. New York, N.Y.: John Wiley, 1959. p. 240-242.

9. **Gumpertz, M., and S.G. Pantula**: A Simple Approach to Inference in Random Coefficient Models. *American Statistician* 43:203-210 (1989).

10. **SAS Institute**: *SAS/STAT Software: Changes and Enhancements through Release 6.12.* Cary, N.C.: SAS Institute, 1997. pp. 531-570.

11. **Fischbach, T., R. Song, and S. Shulman**: Some Statistical Procedures For Analytical Method Accuracy Tests and Estimation. *Am. Ind. Hyg. Assoc. J.* 57:440-451 (1996).

12. **Shulman, S., and J. Smith**: "Estimation of Bias and Variability of Continuous Gas Monitor Data: Application to Carbon Monoxide Monitor Accuracy." Poster presented at the Joint Statistical Meetings, Dallas, Tex., 1998.

13. **Box, G.E.P.**: The Effects of Errors in Factor Levels and Experimental Design. *Bulletin of International Statistical Institute* 38:343-344 (1961).

14. **Searle, S.R.**: *Linear Models*. New York, N.Y.: John Wiley, 1971. p. 413.

15. **Giesbrecht, F.G., and J.C. Burns**: Two-Stage Analysis Based on a Mixed Model: Large Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics* 41:477-486 (1985).

16. **Miller Jr., R.G.**: *Simultaneous Statistical Inference*. 2nd ed. New York, N.Y.: Springer-Verlag, 1980. p. 8.

17. Shulman, S. A. and J. P. Smith. " **Description of a Method for Estimation of Bias and Variability of Continuous Gas Monitor Data:** Application to Carbon Monoxide Monitor Accuracy, " Submitted to AIHAJ.

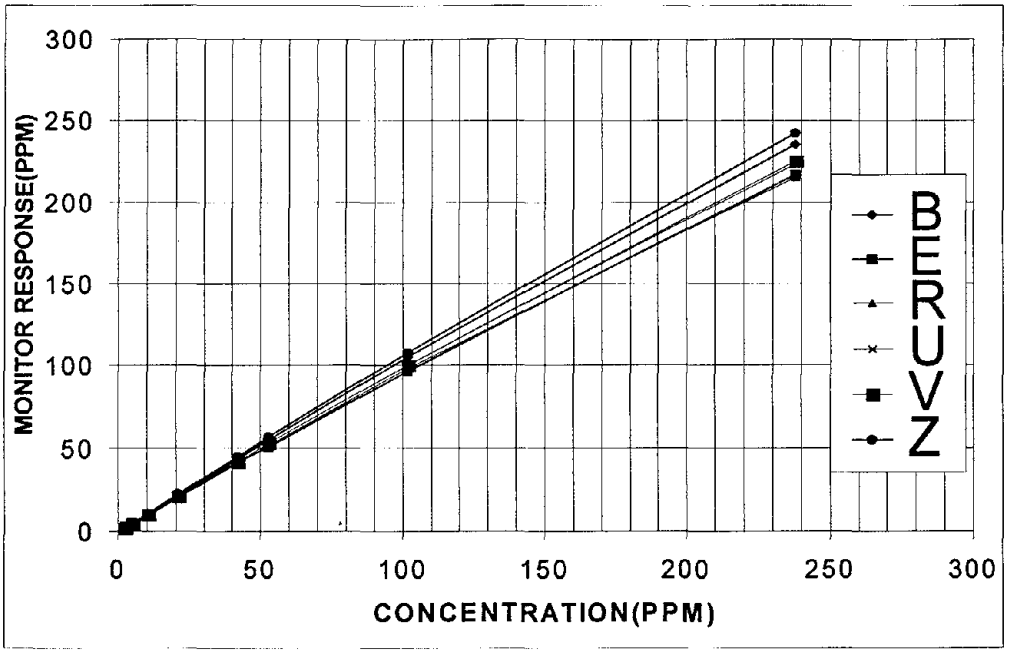
TABLE I: Accuracy Estimates for Example Carbon Monoxide Monitor Data.

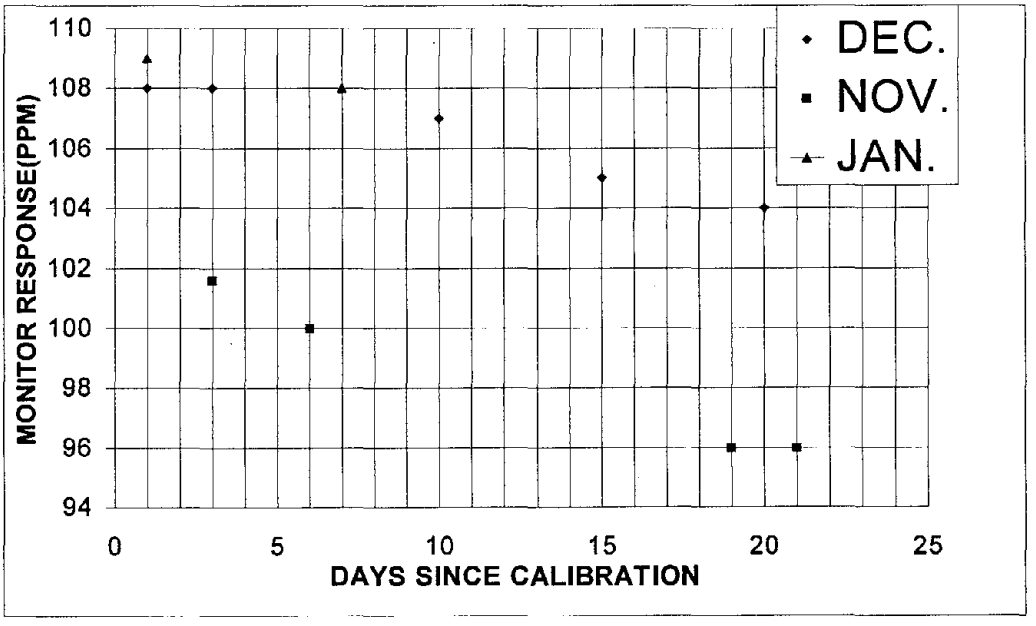
Conc	Bias Est.	$\hat{S}_{rT,tr}$	Accur Est.	Accur (ppm)	Deg Fr(r2)	Deg Fr(r1)	Conf.Lim. Bias	Conf.Lim. $\hat{S}_{rT,t}$	Up 95% Accur	Up 95% Accur. (ppm)
10 ppm	-0.131	0.135	0.353	±3.53	3.00	2.13	-0.420	0.503	1.25	±12.5
30 ppm	-0.0186	0.0893	0.180	±5.40	3.51	2.45	-0.189	0.286	0.669	±20.1
50 ppm	-0.0025	0.0819	0.161	±8.05	3.94	2.63	-0.149	0.238	0.549	±27.5

FIGURE CAPTION LIST

Figure 1. Response Curves for Each of the Six Monitors, for One Day's Determinations.

Figure 2. Monitor Response for One of the Monitors as a Function of Day Since Calibration.


MONITOR RESPONSE(PPM)
300
250
200
150
100
50
0
0
50
100
150
200
250
300
CONCENTRATION(PPM)
BERUVZ


MONITOR RESPONSE(PPM)
110
108
106
104
102
100
98
96
94
0
5
10
15
20
25
DAYS SINCE CALIBRATION
DEC.
NOV.
JAN.