AIHA Journal

ISSN: 1542-8117 (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaih20

A Method for Estimation of Bias and Variability of Continuous Gas Monitor Data: Application to Carbon Monoxide Monitor Accuracy

Stanley A. Shulman & Jerome P. Smith

To cite this article: Stanley A. Shulman & Jerome P. Smith (2002) A Method for Estimation of Bias and Variability of Continuous Gas Monitor Data: Application to Carbon Monoxide Monitor Accuracy, AIHA Journal, 63:5, 559-566, DOI: 10.1080/15428110208984740

To link to this article: https://doi.org/10.1080/15428110208984740

AIHA Journal 63:559–566 (2002) Ms. #272

A Method for Estimation of Bias and Variability of Continuous Gas Monitor Data: Application to Carbon Monoxide Monitor Accuracy

AUTHORS
Stanley A. Shulman
Jerome P. Smith

National Institute for Occupational Safety and Health, Division of Applied Research and Technology, 4676 Columbia Parkway—R3, Cincinnati, OH 45226

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. Although the most general statistical software required for these tasks is not available on a spreadsheet, when the same number of dates in a calibration period are evaluated for each monitor unit, the calculations can be done on a spreadsheet. An example of such calculations, together with the formulas needed for their implementation, is provided. In addition, the methods can be extended by use of appropriate statistical models and software to evaluate monitor trends within calibration periods, as well as consider the effects of other variables, such as humidity and temperature, on monitor variability and bias.

Keywords: accuracy, bias, evaluation of gas monitors, precision

Direct-reading monitors are available for a wide range of workplace contaminants.[1] Different versions of these instruments may be used for area monitoring in one location, for area monitoring in different locations, and for personal monitoring. Data obtained from them 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 their usefulness and convenience in assessing exposure.

Because 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 these instruments 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 do not allow calculation of overall accuracy of the instrumental data.

This article is intended to introduce the concept of statistical modeling of this kind of data, but not to give the statistical details. The detailed description of the statistical procedures is contained in another document.[4] Statistical modeling of such data will allow estimation of accuracy, much like accuracy evaluation of analytical methods for gas and particulate matter.[5,6] Because decisions concerning use, development, and effectiveness of controls depend on data accuracy, this kind of evaluation is important.

This study developed a procedure to do such an accuracy evaluation. The procedure requires that the response of the monitor to target concentrations be modeled as a linear or higher

order (for example, quadratic) function of these concentrations. In the simplest case the response function will be a straight line, although more complex response functions also can be treated. The method used to generate these target concentrations must meet certain requirements so that the true concentration has the target concentration as its mean and individual deviations from the target are small. If necessary the target concentrations can be verified by an independent method.

Statistical analysis of the response function can be done while a number of variables are changed. For instance, the response function of a set of monitor units can be determined repeatedly over a period of time with a fixed calibration interval to determine response stability for this calibration interval. The response curve also can be determined while changing other variables such as temperature, relative humidity, and interferences to determine their affect on instrumental response and accuracy.

Accuracy is defined in terms of bias and variability.[6] Bias is estimated by the ratio of the monitor's response to the target concentration. Variability is expressed as the relative standard deviation (RSD)—the ratio of the standard deviation to the target concentration. The combined effects of bias and of variability are assessed via accuracy, which is computed so that the probability equals 95% that a measurement will fall within A% of the true value. The current criterion for analytical methods developed by the National Institute for Occupational Safety and Health (NIOSH) requires that $A \leq 25\%$.[6] If bias and variability are estimated rather than treated as known, then a second probability corresponding to the chance of obtaining accuracy A associated with the first probability may be estimated. The work presented in this article enables the user to obtain an estimate of the value of A, such that 95% of individual measurements are within A% of the target value. Also, this work enables the user to obtain an upper 95% confidence limit on the value of A. Note that both bias and precision are defined relative to the target concentration to which the monitors are exposed. The concentration generated must be checked by an independent method to determine that, on average, the targets are approximately attained. This is important. If the generation process has a systematic bias, that bias cannot be separated from the monitor bias, because no comparison to an independent method is provided for in these tests.

In the example used here, monitor response was measured as a function of concentration over an extended time period. A number of identical monitors for carbon monoxide were calibrated simultaneously at fixed intervals during this period and evaluated simultaneously at various times between calibrations. The main purpose was to determine whether the calibration procedure and frequency were adequate to fulfill the accuracy requirement. Answering these questions required statistical modeling of the data so that differences in response with time and with monitor could be identified.

MODEL FOR MONITOR RESPONSE

The model given below describes a procedure for estimating the accuracy of the measurements taken by the monitor over an extended time period. The accuracy is estimated for randomly chosen monitors, which are calibrated at the beginning of each calibration period, and subsequently evaluated at randomly chosen times in the period. Monitors will be assessed by fitting a response curve for each trial, in each of which the monitors are exposed to several different concentrations of analyte. A related example is given in Vecchia[7] of assessment of the variability over time of analytical instrument calibration lines. The purpose of evaluating monitor data is the assessment of the adequacy of the calibration period. If accuracy is within predetermined limits, then this calibration period will be adopted.

The monitor response is designated as $y_{pdm,j}$ for a given monitor (m) on day (d) in calibration period (p) at concentration C_j. As an example, the method will be shown for a linear response function; however, the method can be used for polynomial response, too:

$$y_{pdm,j} = \beta_{pdm,cep} + \beta_{pdm,lin}\, C_j + e_{pdm,j} \quad (1)$$

In Equation 1 the subscript "cep" denotes the intercept of the response curve and "lin" denotes the linear parameter of the response curve. Because there are J concentrations, y_{pdm} may be written as a (J by 1)-dimensional vector of responses at the target concentrations. For example, if there are five concentrations then

$$y_{pdm} = \begin{vmatrix} y_{pdm,1} \\ y_{pdm,2} \\ y_{pdm,3} \\ y_{pdm,4} \\ y_{pdm,5} \end{vmatrix} = \begin{vmatrix} 1 & C_1 \\ 1 & C_2 \\ 1 & C_3 \\ 1 & C_4 \\ 1 & C_5 \end{vmatrix} \begin{vmatrix} \beta_{pdm,cep} \\ \beta_{pdm,lin} \end{vmatrix} + \begin{vmatrix} e_{pdm,1} \\ e_{pdm,2} \\ e_{pdm,3} \\ e_{pdm,4} \\ e_{pdm,5} \end{vmatrix} \quad (2)$$

or $y_{pdm} = X\,\beta_{pdm} + e_{pdm}$, where X is a matrix containing the concentrations. In the above expression,

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

$$\beta_{pdm} = \begin{vmatrix} \beta_{cep} \\ \beta_{lin} \end{vmatrix} + \begin{vmatrix} a_{p,cep} \\ a_{p,lin} \end{vmatrix} + \begin{vmatrix} b_{pd,cep} \\ b_{pd,lin} \end{vmatrix} + \begin{vmatrix} c_{m,cep} \\ c_{m,lin} \end{vmatrix} + \begin{vmatrix} ac_{pm,cep} \\ ac_{pm,lin} \end{vmatrix} + \begin{vmatrix} abc_{pdm,cep} \\ abc_{pdm,lin} \end{vmatrix}. \quad (3)$$

Each of the addends in Equation 3 is a vector that represents variability of the components of β_{pdm} over periods (p), days (d), or monitors (m) or combinations of these. 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. 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 in Equation 2 for the linear response curve model. In this model the chosen monitors, days within calibration period, and chosen calibration periods are viewed as random samples from a larger population. In some instances, for example, if there are trends within each calibration period, this assumption may be violated. A solution is to use in the statistical analysis only those days in each period that are closer together in time, thereby eliminating the effect of trend.

The same form as Equations 2 and 3 could be used to describe a polynomial relation. For instance, by adding a third column of squared concentrations to the X matrix and a quadratic parameter to the b vector in the third row, the Model 2 can be used to describe a quadratic response.

If the parameters of interest were means, rather than the parameters associated with a curve, each of the addends in Equation 3 would be one dimensional. For instance, suppose that m monitors were used to measure a property of an industrial product, on day d within period p. If the interest were in the variability of the product over monitors, between periods, and over days within period, then Model 3 might be appropriate, except that each component would be one dimensional. This one-dimensional form is a form that often arises in quality control.

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 measurement errors are instantaneous and represent the random noise associated with the monitor determinations. On the other hand, the variability of the components of β_{pdm} is of a different nature, whether between calibration periods, between days in period, or between monitors. Also, it is best if the same concentrations are used at each evaluation. Because the concentrations are set by the experimenter, this is not an unreasonable aim. Also, if monitors are simultaneously evaluated in a chamber, then it is assumed that deviations of the generated concentrations from the target concentrations are small compared with the other sources of error, which is reasonable in a well-controlled chamber.

COMPUTER CODE FOR EVALUATING MONITOR ACCURACY

A SAS[8] program has been written that fits a linear or quadratic curve for each monitor evaluation, and these results can then be used to obtain estimates of the accuracy A, and the upper 95% confidence limit for A. The program assumes that all fitted curves are fitted at the same concentration values. This program is appropriate for situations in which randomly chosen monitors of the same kind are periodically recalibrated at about the same time and are evaluated several times between calibrations. This evaluation may be done either simultaneously or one at a time. The required parameter estimates are obtained in two steps. The program first uses PROC REG to fit the linear or 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. A reason to prefer the two-step procedure is that, in the authors' experience, there may sometimes be convergence problems in the larger data set associated with the one-step procedure. The result of "convergence problems" is that the computer program cannot fit the statistical model. The two-step procedure is also consistent with the procedure for obtaining estimated variances for balanced data that is derived elsewhere.[4] The term "balanced" is used to describe the special situation when all monitor units are evaluated the same number of days in each calibration period. Such data are called balanced with respect to calibration period, days in period, and monitor.

For balanced data the accuracy computations can then be done in Excel. An example of these computations is given in the Appendix 1. If the data are nearly balanced, so that they could be made balanced by, for example, averaging or deleting some of the days' evaluations, then the spreadsheet method shown in the appendix can also be used. Although the steps used in the spreadsheet are documented in the appendix, they are numerous and complicated. The spreadsheet is set up so that if the mean squares and degrees of freedom are provided in step F (see appendix), then the sheet will calculate the precision, bias, and accuracy, and their 95% confidence limits. The spreadsheet is available from the authors.

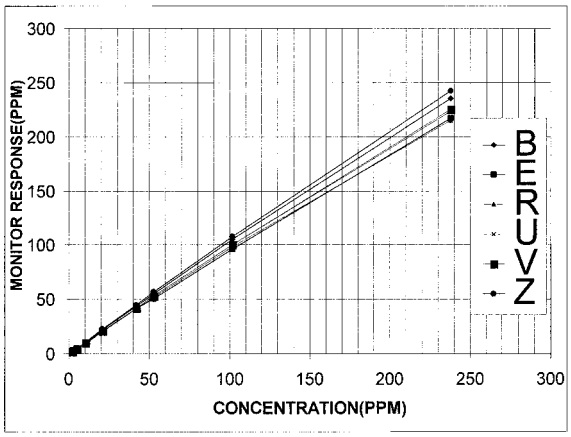

FIGURE 1. Response curves for each of the six monitors, for one day's determinations

EXAMPLE USE OF PROCEDURE—CARBON MONOXIDE MONITORS

An accuracy evaluation of personal data logging monitors for carbon monoxide was done.[9] 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 0 span gas and to 250 ppm CO. The calibration mixture and equipment were obtained from the manufacturer and were used according to manufacturer procedures. The monitor was forced to have its readout agree at the 0 and 250 ppm concentrations. The assumption is then made that the response 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.

The monitors were placed in a chamber where they were simultaneously exposed to varying concentrations of CO during each exposure trial. Between two and six of these exposure trials were conducted in each of the three calibration periods. 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 the readings stabilized. The flow controllers were periodically recalibrated, and the concentration in the chamber was checked by an independent method (Fourier transform infrared). The atmosphere in the chamber was under low and well-mixed flow so that all parts of the chamber would be exposed to the same concentration at close to the same time. Humidity was evaluated in another part of study, and its effect was not found to be statistically significant.[9] Humidity should have been similar for experiments in the same calibration period.

Twelve days of data collected over 3 months were used in this study. The number of days varied in each month. 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

TABLE I. Accuracy Estimates for Example Carbon Monoxide Monitor Data

Concentration	Bias Est.	RSD Relative to Target Value	Accur Est.	Accuracy (ppm)	Conf. Lim. Bias	Conf. Lim. RSD	Up 95% Accuracy	Up 95% Accuracy (ppm)
10 ppm	−0.131	0.135	0.353	±3.53	−0.420	0.503	1.25	±12.5
30 ppm	−0.0186	0.0893	0.180	±5.40	−0.189	0.286	0.669	±20.1
50 ppm	−0.0025	0.0819	0.161	±8.05	−0.149	0.238	0.549	±27.5

1 all determinations made by the same monitor have been connected. The figure indicates the small curvature of the response lines and also indicates 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 were expected to be correlated with each other, though the evidence was that the correlation was small. (The correlation was estimated from the correlation of the residuals of the individually fitted response curves of the monitors evaluated simultaneously.) Plots of the data indicated no apparent systematic biases among the monitors.

Because each of the 12 days of data contributed six sets of parameter estimates (one for each of the six monitors), there were 72 fits of a quadratic response curve. For the 3 months of the example data, the largest component of variability was among months. Because this was the largest component, the degrees of freedom of the estimated total variance was closest to the degrees of freedom of this largest component—between 3.00 and 3.94 for concentrations between 10 and 50 ppm. (When the mean squares in an analysis of variance are statistically independent and follow chi-square distributions, Satterthwaite's approximation can be used to estimate the degrees of freedom[10] of a linear combination of mean squares, as is required here for the total variance. The estimated degrees of freedom is not necessarily an integer, and will be closest to the degrees of freedom of the dominant mean square.)

For these data the estimated average parameter values and total variance matrix were as follows: $(-1.785\ \ 1.052\ \ -0.000379)'$, where -1.785 was the average value for the intercept estimate, 1.052 was the average value for the linear parameter estimate, and -0.000379 was the average value for the quadratic parameter estimate. Thus, the fitted average response curve was:

$$y_{pdm,j} = (-1.785 + 1.052C_j - 0.000379\ C_j + f_{pdm,j})$$

where the error term $f_{pdm,j}$ has components of variance due to calibration period, day within calibration period, monitor, monitor by calibration period, and monitor by day within calibration period. (These components also appear in Equation 3.) The largest component of the total variance was due to calibration periods, constituting over 60% of the total variance at 10, 30, and 50 ppm. Of the remaining 40%, the size of the components varied by the concentration. For instance, at 10 ppm the day within calibration period was the next largest component, but at 50 ppm, monitor variability was second largest.

Because monitor performance at low concentrations is of interest, the accuracy estimates and individual 95% confidence bands at low concentrations (10, 30, and 50 ppm) are shown in Table I, which contains the following information: estimated bias; estimated RSD; estimated accuracy as fraction of concentration; estimated accuracy in parts per million; lower 2.5% confidence limit on bias; upper 97.5% confidence limit for RSD; upper 95% confidence limit for accuracy as fraction of concentration; and upper 95% confidence limit on accuracy in parts per million.

Notice that in Table I the confidence limit for the bias is the 2.5% limit, and that for RSD is the 97.5% limit. (When the bias estimate is negative, the 2.5% limit should be used; when it is positive, the 97.5% limit should be used. In either case, the aim is to get the most extreme value for the particular situation.[6]) When substituted into the accuracy equation,[11] an upper 95% confidence limit for accuracy was obtained. The accuracy estimates in the fourth column of the table are similar (although somewhat larger) to the manufacturer's specifications. For instance, the accuracy estimate at 10 ppm was 0.353, which means that it is estimated that 95% of the individual measurements taken at 10 ppm were within $\pm(0.353)(10)$, or were within ±3.53 ppm of 10 ppm, or were between 6.47 and 13.53 ppm. (Notice that the value 3.53 ppm is given in column 5 of Table I.). The accuracy 95% upper confidence limits in the last column of the table (for instance, ±12.5ppm at target value of 10 ppm) indicate that the accuracy based on the evaluation experiment was 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, the manufacturer's specifications were most likely based on single measurements made on the same day that the monitors were calibrated. For the estimates presented here the measurements were made days after calibration. The monitors tended to underestimate the concentration, though the size of the underestimate varied with the concentration. The upper limits on accuracy were much bigger than the accuracy estimate because the main source of variability, calibration period (accounting for more than half the variance of the linear parameter), had few degrees of freedom—2, or because the true accuracies were 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 of bias and RSD, which determine both the accuracy estimate and the confidence limit on accuracy. If the accuracy estimate is higher, then the upper confidence limit on accuracy will be higher.

Within calibration period and regardless of the concentration, the monitors tended to produce 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 3 months of data. Figure 2 also indicates the amount of variability among the three calibration periods. The main variability appeared to be between November and the other two months.

In the model used here the days within calibration period were treated as a random sample over which the response curve parameters varied. This may not have been an appropriate assumption in this situation. The statistical analysis did indicate that at the 30 and 50 ppm concentrations, there was statistically significant decrease with day since calibration. When this was allowed for in the statistical model, the upper confidence limits for accuracy improved somewhat for analyses carried out at about 10 days after calibration (the average of the days since calibration for which

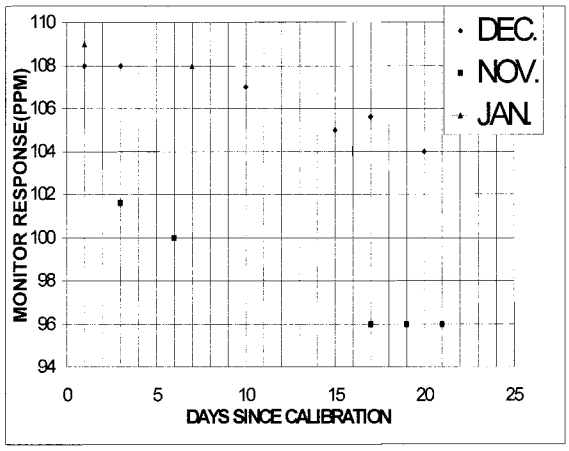

FIGURE 2. Monitor response for one of the monitors as a function of days since calibration

there were data). For 10, 30, and 50 ppm the new upper accuracy limits were 12.3, 17.8, and 22.8 ppm, which were improvements over the 12.5, 20.1, and 27.5 ppm shown in Table I. Improvement was to be expected, as by allowing for dependence on day since calibration, the variances used in accuracy calculations were reduced.

In summary, bias tended 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 exceeded the 2-ppm accuracy limit provided by the manufacturer, and exceeded 12 ppm when the variability of the estimates was taken into account. Accuracy limits were wide, because more calibration periods should have been studied. The effect of having only three calibration periods is, first, that there were few degrees of freedom for calibration periods, and, second, that it is unknown whether the low results for November should be regarded as outlying results. Had the variance estimates been the same but with 10 calibration periods, the upper 95% limit on accuracy at 10 ppm would have been approximately ±7 ppm, much less than the ±12.5 ppm shown in Table I. This reduction demonstrates the importance of the number of calibration periods. With the November data removed, at 10 ppm the total variance was reduced by about 75%, and the upper limit on accuracy was about ±5 ppm. If there were more than two calibration periods remaining after deletion of the November data, the upper accuracy limits would be even lower. The upper limits shown here are much smaller than those given in Table I, but because of uncertainty about the reason that the November data were low, it seems best not to exclude them. Therefore, the authors repeat that accuracy limits were wide because more calibration periods should have been studied. Therefore, this evaluation should not be regarded as a complete evaluation of these monitors.

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 given in this article. The model can be extended to examine these effects by including terms in Model 3 that are functions of these additional variables. The SAS computer code can be modified to include these new variables.

CONCLUSIONS

A method was presented for the evaluation of the bias, variability, and accuracy of gas monitors at many different concentrations by using estimates from a single statistical model.

Although application of this approach to carbon monoxide monitor data suggested large variability between calibration periods, insufficient periods were studied to be certain of this finding. Also, because the data indicated a decrease in monitor determinations within calibration period, it would be beneficial to calibrate monitors more frequently than once a month. The mathematics involved in using the model is somewhat complicated and is presented in more detail elsewhere.[4] However, for balanced data, calculations can be done on a spreadsheet, details of which are presented in the appendix to this article. This should enhance the applicability of the method. Also, the method can be generalized to evaluate the effect of additional variables, such as temperature and humidity.

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 preparation of the manuscript.

REFERENCES

1. **Cohen, B., and S. Hering (eds.):** *Air Sampling Instruments for Evaluation of Atmospheric Contaminants,* 8th ed. Cincinnati, Ohio: American Conference of Governmental Industrial Hygienists, 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. **Shulman, S.A., and J.P. Smith:** *Estimation of Bias and Variability of Continuous Gas Monitor Data: Application to Carbon Monoxide Monitor Accuracy.* (PB2002104394). Springfield, VA: National Technical Information Service, 2001.
5. **Busch, K.A., and D.G. Taylor:** Statistical protocol for the NIOSH validation tests. In G. Choudhary, editor, *Chemical Hazards in the Workplace—Measurement and Control* (Symposium Series 149). Washington, D.C.: American Chemical Society, 1981. pp. 503–517.
6. **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 publication no. 95–117). Washington, D.C.: Government Printing Office, 1995.
7. **Vecchia, D.F., H. Iyer, and P.L. Chapman:** Calibration with randomly changing standard curves. *Technometrics 31*(1):83–90 (1989).
8. **SAS Institute:** *SAS/STAT Software: Changes and Enhancements Through Release 6.12.* Cary, N.C.: SAS Institute, 1997.

9. **Smith, J.P., and S. Shulman:** Evaluation of a personal data logging monitor for carbon monoxide. *Appl. Occup. Environ. Hyg. 9*:418–427 (1994).

10. **Searle, S.R.:** *Linear Models.* New York: John Wiley, 1971.

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 Joint Statistical Meetings, Dallas, Tex., 1998.

APPENDIX

A DATA

Period	Date	Monitor	Standard	Response
1	1	1	1	-2
1	1	1	10	17
1	1	1	20	36
1	1	2	1	-0.5
1	1	2	10	17
1	1	2	20	28
1	13	1	1	-0.5
1	13	1	10	15
1	13	1	20	29
1	13	2	1	-1
1	13	2	10	16
1	13	2	20	29
2	3	1	1	1
2	3	1	10	11
2	3	1	20	22
2	3	2	1	0.5
2	3	2	10	9
2	3	2	20	16
2	15	1	1	3.5
2	15	1	10	17
2	15	1	20	31
2	15	2	1	0.75
2	15	2	10	11
2	15	2	20	18

B ESTIMATES

Period	Date	Monitor	Slopes	Intercept	Sum	Residual MS
1		1	2.00	-3.65	-1.65	0.67
1	1	2	1.49	-0.59	0.90	10.66
1	13	1	1.55	-1.51	0.04	1.55
1	13	2	1.57	-1.60	-0.02	5.18
2	3	1	1.11	-0.09	1.02	0.00
2	3	2	0.81	0.09	0.91	0.89
2	15	1	1.45	2.22	3.67	0.15
2	15	2	0.90	0.57	1.48	2.88
		Averages:	1.36	-0.57	0.79	2.75

Estimated Bias= 1.30

C SLOPES

Period	Monitor Unit 1	Monitor Unit 2
1	2.00	1.49
	1.55	1.57
2	1.11	0.81
	1.45	0.90

Period	Date 1	Date 2
1	2.00	1.55
	1.49	1.57
2	1.11	1.45
	0.81	0.90

ANOVA

Source	SS	df
Period	0.69	1.00
Monitor	0.22	1.00
Period* Monitor	0.02	1.00
Total	1.09	7.00

ANOVA

Source	SS	df
Period	0.69	1.00
Date	0.00	1.00
Inter-action	0.08	1.00
Date (Period)	0.08	2.00

D INTERCEPTS

	Monitor Unit 1	Monitor Unit 2
1	-3.65	-0.59
	-1.51	-1.60
2	-0.09	0.09
	2.22	0.57

	Date 1	Date 2
1	-3.65	-1.51
	-0.59	-1.60
2	-0.09	2.22
	0.09	0.57

ANOVA

Source	SS	df
Period	12.87	1.00
Monitors	0.29	1.00
Period* Monitor	2.47	1.00
Total	21.18	7.00

ANOVA

Source	SS	df
Period	12.87	1.00
Date	1.92	1.00
Interaction	0.35	1.00
Date (Period)	2.26	2.00

E SUM(SLOPE+INTERCEPT)

	Monitor Unit 1	Monitor Unit 2
1	-1.65	0.90
	0.04	-0.02
2	1.02	0.91
	3.67	1.48

	Date 1	Date 2
1	-1.65	0.04
	0.90	-0.02
2	1.02	3.67
	0.91	1.48

ANOVA

Source	SS	df
Period	7.61	1.00
Monitors	0.00	1.00
Period* Monitor	2.87	1.00
Total	16.0	7.00

ANOVA

Source	SS	df
Sample	7.61	1.00
Date	1.98	1.00
Interaction	0.76	1.00
Date (Period)	2.74	2.00

F **Mean Squares**

	Degrees of Freedom(df)	Number	From D Intercept	From C Slope	From E Sum	Covariance at x=10	Mean Square at x=10
Period	1.00	2.00	12.87	0.69	7.61	-2.97	22.12
Monitors	1.00	2.00	0.29	0.22	0.00	-0.25	16.95
Period* Monitor	1.00	2.00	2.47	0.02	2.87	0.20	7.93
Date(Period)	2.00	2.00	1.13	0.04	1.37	0.10	7.12
Residual	2.00	2.00	1.65	0.04	1.39	-0.15	2.93

G **VARIANCE COMPONENTS**

	Intercept	Slope	Sum	Covariance	Variance Component at X=10		Variance Components at Mean of X=10
Period	2.73	0.17	1.19	-0.85	2.50		1.25
Monitors	-0.55	0.05	-0.72	-0.11	2.25		1.13
Period* Monitor	0.41	-0.01	0.74	0.17	2.50		0.63
Date(Period)	-0.26	0.00	-0.01	0.12	2.10		0.52
Residual	1.65	0.04	1.39	-0.15	2.93		0.37
Total	3.98	0.25	2.59	-0.82	12.28	=Total Variance of Future Measurement at X=10	Variance of Mean at x=10 3.89

H **DEGREES OF FREEDOM**

	For Variance at x=10				For Variance(Mean) at x=10		
	Mean Square, x=10	Multiplier	Mean Square* Multiplier	Component Degrees of Freedom	Multiplier of Mean Square, x=10	Mean Square* Multiplier	Degrees of Freedom,
Period	22.12	0.25	5.53	30.57	0.13	2.76	7.64
Monitors	16.95	0.25	4.24	17.96	0.13	2.12	4.49
Period* Monitor	7.93	0.00	0.00	0.00	-0.13	-0.99	0.98
Date(Period)	7.12	0.25	1.78	1.59	0.00	0.00	0.00
Residual	2.93	0.25	0.73	0.27	0.00	0.00	0.00
				50.38			13.12

Degrees Freedom, Individual Prediction, x=10 2.99

Degrees of Freedom, for Mean at x=10 1.15

Variance (Individual) Prediction, x=10 12.28

Variance of Mean at x=10 3.89

I **CONFIDENCE LIMITS**

Estimated RSD 0.35

RSD< 1.31 (Uses linear interpolation)

J **ACCURACY**

95% Accuracy< 4.65

Estimated Bias< 1.30

If Bias >0 Bias< 3.55 (Uses linear interpolation)

If Bias <0 Bias> -1.21

Notes: Contribution of variability about fitted lines and correlation among determinations made simultaneously considered negligible. Both can be taken into account in a fuller analysis. See notes to individual tables (A through J, following) for explanations of components of the table.

A. The data, shown by calibration period, date in period, monitor unit, standard concentration (x), and monitor response (y). Data are were made up for the calculations.

B. Estimates for intercepts, slopes, and sum of these given for each of eight lines, as well as the residual mean square. Produced via "regression" method in "Data Analysis" in "Tools." (These are ordinary least squares estimates. If weighted least squares estimates are needed, alterations must be made to the spreadsheet.) Averages were produced for each of these estimates. Average (ratio) bias at x = 10 is (−0.57 + 1.36*10)/10 = 1.3, where −0.57 was the average intercept value and 1.36 was the average slope.

C, D, E. Use "ANOVA: Two Factor with Replication" from the "Tools" package. Thereby the mean squares and their degrees of freedom were obtained, for calibration period, date in period, monitor, monitor by period interaction, and the residual. These were computed for intercept, slope, and the sum of these. For instance, in D the mean square of periods for the intercept was 12.87, with 1 degree of freedom.

F. The residual mean squares were computed as: (total sum of squares − sum of all component sums of squares), divided by the degrees of freedom. The mean squares from D, E, F were used here, as indicated in the table. For instance, the 12.87 was the period mean square for the intercept. Its degrees of freedom, 1, is two columns to the left, and the column "num" is the number of periods. The column labeled "covar" has, for each component, the value 0.5(sum mean square − intercept mean square-slope mean square), which for period is: 0.5(7.61 − 12.87 − 0.69) = −2.97. The next column contains the mean

squares for x = 10, computed as: [intercept MS + 10(covariance)] + 10[covariance + 10(slope MS)]. For the period component this was [12.87 + 10(−2.97)] + 10[−2.97 + 10(0.69)] = 22.12.

G. Mean squares from F were used to compute the variance components. Var(period*monitor) = [MS(per*mon) − MS(residual)]/num(dates in period). For intercept the period*monitor variance is Var(period*monitor) = (2.47 − 1.65)/2 = 0.41, Var(monitor) = [MS(mon) − MS(per*mon)]/[num(dates in period)*num(periods)], Var(date in period) = [MS(date in period) − MS(residual)]/num(monitors), Var(period) = [MS(per) − MS(date in per) − MS(per*mon) + MS(residual)]/[num(dates in period)*num(mon)].

Covariance components were caculated as 0.5(sum var comp − intercept var comp-slope var comp). For period, this was: 0.5(1.19 − 2.73 − 0.17) = −0.85. These components were combined as follows to obtain the component at x = 10: [intercept comp + 10(covariance comp)] + 10[covariance comp + 10(slope comp)]. For period this was: [2.73 + 10(−0.85)] + 10[−0.85 + 10(0.17)] = 2.5. The sum of variance components gives total variance of a future measurement at x = 10, or 12.28. The different components should be examined to see which make large contributions to the total. In the calculation of the variance of the average measurement at x = 10, each of the above variance components is divided by the following numbers: var(per) by num(per); var(mon) by num(mon); var(per*mon) by [num(per)*num[mon)]; var(date in per) by [num(date)*num(per)]; var(res) by number of measurements. For instance, the period variance component for individual measurements at x = 10 is 2.5. Divided by 2 this becomes 1.25 for the mean at x = 10. The different components were summed to obtain the variance of the estimated mean monitor response at x = 10. This variance was 3.89 = (1.25 + 1.13 + 0.63 + 0.52 + 0.37).

H. Degrees of freedom (df) for an individual predicted value at x = 10 requires the following formulas: (#) df = (total var)^2/sum, where sum = {MS(per)/[num(mon)*num(date in per)]}^{2}/df(per) + {MS(mon)/[num(per)*num(date in per)]}^{2}/df(mon) + {MS(per*mon)*{1/num(date in per)−1/[num(date in per)*num(mon)] − 1/[num(date in per)*num(per)]}}^{2}/df(per*mon) + {MS(date in per)*[1/num(mon) − 1/[num(date in per)*num(mon)]]}^{2}/df(date) + {MS(res)*[1 + 1/[num(mon)*num(date in per)] − 1/num(date in per) − 1/num(mon)]}^{2}/df(res).

In the column labeled "multiplier," the appropriate coefficient for that MS appears. For instance, for period the multiplier is 1/[num(mon)*num(date in per)], which is 0.25 here. [(0.25)22.12]^2/1 = 30.57, which appears in the column labeled "deg fr." (22.12 was the mean square for periods at x = 10, from F.) The total of the column labeled degrees of freedom was sum = 50.38, which was entered into the df formula (# above) to obtain df = (12.28)^2/50.38, or 2.99 degrees of freedom associated with the total variance (12.28) of a future measurement at x = 10, calculated in G.

The degrees of freedom calculations for the mean value at x = 10 require the following modifications of the above formulas for the multipliers: period "multip" for mean = individual measurement multi/number(periods), monitor "multip" for mean = individual measurement multi/number(mon), per*mon "multip" for mean = −1/[num(date)*num(mon)*num(per)], date in period "multip" for mean = 0, residual "multip" for mean = 0.

When the multipliers were used to compute the denominator "Sum" in (#) above, the result was 13.12. Because the estimated variance for the estimated mean at x = 10 was 3.89 (from G), the degrees of freedom was 3.89(3.89)/13.12, or 1.15.

I. The estimated relative standard deviation (relative to x = 10) for an individual measurement at x = 10 was (12.28)$^{0.5}$/10 = 0.35. The upper 97.5% confidence limit on the true relative standard deviation was {df (.35)^2 chi_inv(.025,df)}$^{0.5}$, where chi_inv(.025,df) is the 0.025 percentile of the chi-square distribution with df degrees of freedom. In the table above, the chi-square value is obtained by linear interpolation because of fractional degrees of freedom resulting from formula (#). This is done by interpolating between the two successive integral degrees of freedom values between which the estimate from (#) is located. Here df∼2.99 (from H), and the linear interpolation is between chi_inv(.025,2) and chi_inv(.025,3), giving a value of 0.214. The upper limit was [(2.99)(0.35)^2/0.214)]$^{0.5}$, approximately 1.31. For the confidence limit on the bias, the formula used was: estimated bias at 10 + (std dev of mean)* tinv(.975, df_mean). Bias at x = 10 is 1.30 (from B). The required standard deviation was (var. of mean/100)$^{0.5}$, or (3.89/100)$^{0.5}$ (from H), 0.20. Tinv(.975, df_mean) was the 97.5 percentile of the Student's t distribution with df degrees of freedom. Here df was about 1.15 (from H), also by linear interpolation—11.40. Thus, the upper confidence limit was 1.3 + (0.20)11.40∼3.55. The lower confidence limit was obtained analogously.

J. The confidence limits were substituted into the accuracy equation:[6] 1.57(upper limit for RSD) + [(0.39*RSD)^2 + (upper limit bias −1)^2]$^{0.5}$, or 1.57(1.31) + {[.39(1.3)]^2 + (3.55 − 1)^2}$^{0.5}$, or 4.66. With 95% probability, an individual measurement will be within 466% of the true value 95% of the time. Had the bias been negative, the lower confidence limit would have been used in the accuracy equation.