Pergamon

Ann. occup. Hyg., Vol. 45, No. 5, pp. 357–364, 2001
Published by Elsevier Science Ltd on behalf of British Occupational Hygiene Society
Printed in Great Britain.
PII: S0003-4878(00)00071-5
0003-4878/01/$20.00

Definition and Assessment of Sampling and Analytical Accuracy

David L. Bartley

National Institute for Occupational Safety and Health, 4676 Columbia Parkway, Cincinnati, OH 45226, USA

Two independent definitions for quantifying measurement accuracy and two limiting schemes for their assessment are examined in this paper. Gauss' *mean square error MSE* is compared to the *symmetric-range accuracy A*, describing the range of measurements about a measurand. Both measures of accuracy account for systematic error (bias) and imprecision so as to quantify the closeness of estimates to the actual values being measured. Remarkably, it is found that the accuracy functions are closely equivalent for most method applications. Furthermore, details are presented on how to compute confidence limits on measurement accuracy so as to account for error in method evaluation. The confidence limits are qualitatively different in the case that the method undergoes extensive initial evaluation in comparison to a continual re-evaluation at each method application. To this end the statistical theories of tolerance as well as more familiar types of confidence intervals are applied. Published by Elsevier Science Ltd on behalf of British Occupational Hygiene Society. All rights reserved

Keywords: accuracy; uncertainty; performance; confidence; tolerance

INTRODUCTION

Sampling and analytical method accuracy continues to hold an important place in the field of occupational hygiene. Adoption of a uniform approach for characterizing accuracy is a current undertaking within both the International Organization for Standardization (ISO) and the American Society for Testing and Materials (ASTM). Recent applications include standardizing performance testing of diffusive samplers of gases and vapors (ISO, 1999); an EC-sponsored inter-comparison of inhalable aerosol samplers (Kenny *et al.*, 1997; Bartley, 1998); and evaluation of respirable aerosol samplers (Kenny and Bartley, 1995). Further, sampling and analytical accuracy is currently under scrutiny in justifying changes proposed in both silica and coal mine dust compliance standards. In general, method accuracy must be known and documented for legally defensible workplace assessments, for controlling exposure, and for making reasonable decisions resulting from epidemiological research.

Received 10 January 2000; in final form 24 August 2000.
E-mail: dbartley@cdc.gov

Consistent with the description of measurement accuracy as the 'closeness of the agreement between the result of a measurement and a true value of the measurand' (ISO, 1993), various means of quantification are possible. This description indicates that both systematic error or bias (if significant) and random error must somehow be accounted for. The elementary approach is to simply specify a pair of numbers corresponding to bias and imprecision. Often, however, the independent information about bias and imprecision is combined in order to make decisions as to method suitability.

Evidently, some information characterizing the measurement system is not included in such a single function. For example, quantitative knowledge of a method's systematic error may allow improved estimation through calibration. Of course, knowledge of a specific accuracy value does not imply discarding other facts about the measurement system.

This paper covers two quantitative implementations of accuracy. Considered first is the *symmetric-range accuracy A*. This function of bias and imprecision is defined as the symmetric range of a specified fraction (for example, 95%) of measurements about the true concentration.

One way of estimating A is through an initial extensive evaluation, prior to making many measurements

Nomenclature

a	accuracy ratio, A/TRSD
A	symmetric-range accuracy or simply *accuracy*, the range symmetric about a true concentration within which a fraction α of measurements are expected to fall
A_γ	confidence limit on symmetric-range accuracy (at confidence level γ)
bias	method bias
c	population mean of a method's estimates of reference concentration C
$\bar{c}$	sample mean of estimates of reference concentration C
$\hat{c}$	a single estimate of reference concentration C
C	known reference or calibration concentration
n	number of estimates in a sample (from a method's evaluation experiment)
k	linear tolerance interval constant
MSE	mean square error
$\hat{\text{MSE}}$	estimated mean square error
MSE_γ	confidence limit on MSE (at confidence level γ)
$\hat{\text{RSD}}$	estimated relative standard deviation ($s/\bar{c}$)
s	estimated standard deviation in estimates $\hat{c}$
$t_{\gamma,\nu}$	Student-t γ-quantile at ν degrees of freedom
TRSD	true relative standard deviation (σ/C)
$\hat{\text{TRSD}}$	TRSD estimate
u_α	unit normal α-quantile
x	population mean of a method's estimates of unknown concentration X
$\hat{x}$	a single estimate of unknown concentration X
X	unknown concentration (to be measured by method)
$\hat{X}$	a single calibrated estimate of unknown concentration X

Greek

α	measurement confidence level (e.g. 95%)
γ	system evaluation confidence level (e.g. 95%)
δ	bias ratio, bias/TRSD
λ	chi-square or Student-t noncentrality parameters
ν	number of degrees of freedom in evaluation experiment ($n-1$)
ν_{eff}	effective degrees of freedom in chi-square approximations
σ	standard deviation in estimates $\hat{c}$
σ_x	standard deviation in estimates $\hat{x}$
σ_γ	confidence limit on σ (at confidence level γ)
$\chi^2_{1-\gamma,\nu}$	chi-square $(1-\gamma)$-quantile at ν degrees of freedom

without re-evaluation, while taking precautions that the method remains stable. Such an application is familiar in industrial hygiene, as most assessments of workplace concentrations of gases, vapors or aerosols are so made. Accuracy of a method may then be characterized by computing a confidence limit on the accuracy A, accounting for error in the method evaluation. In other words, there are two confidences: in the evaluation and in the subsequent measurements.

This definition of accuracy and its assessment are related to the statistical theory of tolerance limits investigated in the 1940s. In fact, in the case of system calibration (to minimize bias) using the evaluation results, the classical tolerance limits are shown to result. More general tolerance intervals result when bias is not minimized.

An alternative accuracy function, the *mean square error* (MSE, proposed by Gauss, 1823) is also discussed. A useful result of this paper is that in many cases of practical interest, MSE and the symmetric-range accuracy A are found to be closely equivalent. Details on how to compute confidence limits on MSE are presented.

An entirely different application, or procedure for use, of a measurement system is then considered. Namely, instead of a single evaluation, the system is re-evaluated at each measurement. Instead of tolerance intervals, confidence limits can be computed on

the measurand by means of a Student-t distribution. Accuracy can then be given as a running average of system performance.

CLASSICAL TOLERANCE INTERVALS AND SYMMETRIC-RANGE ACCURACY

Linear tolerance intervals

Suppose a method is evaluated by taking n measurements $\hat{c}$ (assumed normally distributed about unknown mean c with variance σ^2) of known reference concentration C. Then sample estimates, such as the mean $\bar{c}$ and variance s^2, are easily computed. The classical symmetric tolerance interval (Wilks, 1941, 1942; Wald, 1942, 1943; Hald, 1952) is a special type of confidence interval $[\bar{c}-ks, \bar{c}+ks]$, linear in s about $\bar{c}$. Given a value α (for example, 95%) and evaluation confidence γ (for example, 95%), a constant k exists, independent of c and variance σ^2, so that the probability (upon repeated system evaluations) that a fraction greater than α of future measurements of C falls within the tolerance interval is equal to γ.

An easily applied algorithm for estimating the constant k has been published by Wald and Wolfowitz (1946). In the large n limit, the following simple expression (Hald, 1952) results:

$$k \rightarrow \sqrt{\frac{\nu}{\chi^2_{1-\gamma,\nu}}} \times u_{(\alpha+1)/2}\left(1+\frac{1}{2n}\right), \qquad (1)$$

where $u_{(\alpha+1)/2}$ denotes the unit normal quantile (for example, $u_{(\alpha+1)/2} = 1.960$ at $\alpha = 95\%$), and $\chi^2_{1-\gamma,\nu}$ is the chi-square $(1-\gamma)$-quantile value at $\nu = n-1$ number of degrees of freedom.

Symmetric-range accuracy

An accuracy criterion (Busch, 1977) in use by the US National Institute for Occupational Safety and Health (Gunderson and Anderson, 1980; NIOSH, 1994; Kennedy *et al.*, 1995) for evaluating measurement methods may be interpreted in terms of a type of tolerance interval, which is often identical to the classical limits above. Define the *symmetric-range accuracy*, referred to henceforth as simply the accuracy A, as the fractional range, symmetric about the true concentration (e.g. C), within which α (for example, 95%) of measurements are expected to fall. The accuracy A is, therefore, an increasing function A[bias, TRSD] of bias magnitude |bias| and the true relative standard deviation TRSD, defined as:

$$\begin{aligned} \text{bias} &\equiv /(c-C)/\ C \\ \text{TRSD} &\equiv \sigma/C, \end{aligned} \qquad (2)$$

both approximated here as independent of C. Though not denoted, the accuracy A also depends on the level α.

The accuracy function A[bias, TRSD] may be computed from its implicit definition given above. As shown in Appendix A, the accuracy A is closely approximated by:

$$A = \begin{cases} u_{(1+\alpha)/2} \times [\text{bias}^2 + \text{TRSD}^2]^{1/2}, & |\text{bias}| < \text{TRSD}/u_\alpha \\ |\text{bias}| + u_\alpha \times \text{TRSD}, & \text{otherwise} \end{cases}. \qquad (3)$$

Again, u_α denotes a unit normal quantile (for example, $u_\alpha = 1.645$ at $\alpha = 95\%$). The accuracy of Eq. (3) is indicated in Fig. 1,.where curves of constant accuracy are plotted in the (bias, TRSD)-plane at $\alpha = 95\%$. As seen in Fig. 1, the curves deviate from those of Eq. (3) only close to the indicated sector boundaries, where the maximum fractional error in A is found to equal ±1%.

The NIOSH accuracy criterion requires, on the basis of a method evaluation, that the confidence limit on the accuracy at confidence level $\gamma = 95\%$ is less than 25%. The range defined by the accuracy confidence limit A_γ is then a tolerance interval in that a fraction greater than α of measurements falls inside the interval at probability γ.

Another way of interpreting accuracy A is in terms of confidence intervals on an unknown concentration X. At confidence γ in the evaluation, a fraction α of estimates $\hat{x}$ results in:

$$\frac{\hat{x}}{1+A_\gamma} < X < \frac{\hat{x}}{1-A_\gamma}, \qquad (4)$$

where the confidence limit A_γ is estimated at the reference concentration C.

Accuracy A and the classical tolerance intervals

Often the classical (that is, linear) tolerance limits are directly related to the accuracy confidence level: suppose the results ($\bar{c}$ and s) of the above evaluation experiment for estimating reference concentration C are used to calibrate the method for future measurements of an unknown concentration X. An example would be a one-time calibration of an aerosol size-selective sampler as to size-dependent sampling efficiency. Again bias and TRSD are assumed to be approximately constant. In other words, suppose that, on measuring an unknown concentration X, the estimates and standard deviations scale with X relative to the calibration concentration C as:

$$x/c = \sigma_x/\sigma = X/C. \qquad (5)$$

Then the scaling assumption implies that the interval $X/C[\bar{c}-ks, \bar{c}+ks]$ contains at least a fraction α of estimates $\hat{x}$ at probability γ (though the concentration X is not known). Now, suppose the method is calibrated via the estimate of the reference concentration

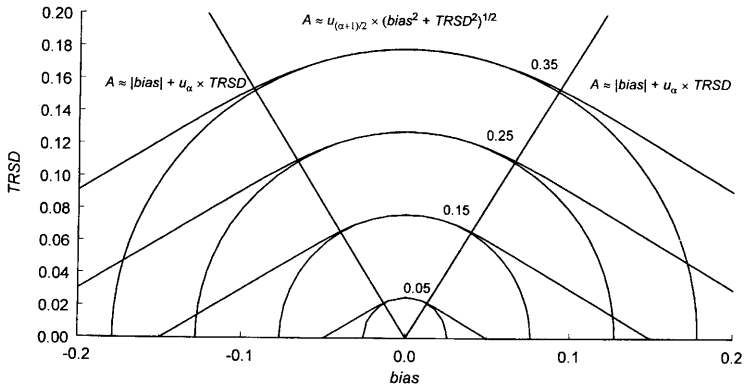

Fig. 1. Curves of constant normalized root mean square error $u_{(\alpha+1)/2}\times\mathrm{MSE}^{1/2}$ (circles) in comparison to accuracy A (quasi-hyperbolas) at $\alpha = 0.95$. Note that at $|\mathrm{bias}|<\mathrm{TRSD}/u_\alpha$ (inside the denoted sector), the two functions are nearly identical. (The values $u_{(\alpha+1)/2} = 1.960$ and $u_\alpha = 1.645$ are unit normal quantiles.)

C, so that the calibrated estimates of X are given by $\hat{X}$:

$$\hat{X} = \frac{C}{\bar{c}}\hat{x}. \tag{6}$$

Then, as shown in Appendix B, the above interval translates to $X[1-ks/\bar{c},\ 1 + ks\bar{c}]$, containing at least a fraction α of estimates $\hat{X}$ at probability γ. As the interval is symmetric about the true (unknown) concentration X, the accuracy A_γ at confidence level γ is given directly by:

$$A_\gamma = k\times s/\bar{c} = k\times\mathrm{R\hat{S}D}\approx \sqrt{\frac{\nu}{\chi^2_{1-\gamma,\nu}}}\times u_{(\alpha+1)/2}\times\left(1+\frac{1}{2n}\right)\times R\hat{S}D, \tag{7}$$

following Eq. (1), where $\mathrm{R\hat{S}D}$ is the estimated relative standard deviation ($s/\bar{c}$). For example, in the important case that $\alpha = 95\%$, $\gamma = 95\%$, and $\nu = 15$,

$$A_{95\%} = 2.9\times\mathrm{R\hat{S}D}. \tag{8}$$

Note that tolerance intervals were applied by Kenny and Lidén (1993) for evaluating the performance of aerosol samplers. Also, the concepts of limit of detection (LOD) and quantitation (LOQ) can be couched in terms of tolerance intervals (ASTM, 2000). Finally, simply asserting that a concentration level is or is not exceeded requires the use of *asymmetric* tolerance intervals.

Uncorrectable bias

As the estimate $\bar{c}$ is almost never equal to c, the calibrated estimate [Eq. (6)] is biased to a degree. No calibration resulting from a finite experiment can correct concentration estimates perfectly. The uncertainty in the calibration is one form of *uncorrectable* bias or systematic error, constant, yet unknown, in each method application subsequent to evaluation. This uncertainty is expressed in the accuracy confidence limit A_γ of Eq. (7) by the factor $[1 + (1/2n)]$. In fact, without this factor, the expression will be recognized by many as a formula commonly used when the bias is known to equal zero prior to the evaluation experiment.

Note that calibration is not always performed as in this example. Some methods call for calibration at each instance of method application. In this case, calibration uncertainty would be a part of the random variability in measurement.

Note further that another common type of uncorrectable bias may be conveniently accounted as an independent component of the method variance. This may occur when the measuring system is sensitive to an environmental parameter (e.g. temperature, aerosol size distribution, etc.), which is impractical to measure during application of the method. Ideally, in this case the method's sensitivity to such an *influence parameter* would be measured. Furthermore, a probability distribution of the parameter would be characterized for the intended application. The effect of its variation then could be treated as part of the random measurement error, rather than bias (ISO, 1993). It is conceivable that other forms of uncorrectable systematic error, such as originating from instability in the sampled material, in the calibration standards, or in the method itself, may be similarly handled as imprecision.

How to account for some types of uncorrectable bias, however, could depend specifically on the application. For example, suppose that a particular method is difficult to implement identically in different labs (owing to differences in lab practices or variability in calibration procedures or standards). Then, acknowledging the existence of interlab bias prior to improving the method, a single lab may be called upon to serve as a primary lab, defining concentrations of an analyte *operationally*.

Uncorrected bias

There exist situations in which bias is not corrected through calibration as above. For example, the distri-

bution of an influence parameter may be difficult to measure, or a bias correction may not be done for historical or other reasons related to method implementation policy. However, even in the case of non-negligible bias, the method accuracy and its confidence limit may be a useful characteristic of the system for judging its utility.

Calculation of confidence limits in this case depends specifically on how the bias is characterized. However, suppose the method is evaluated as above, though without correction through calibration. Several approaches to calculating confidence limits have been attempted:

1. The original NIOSH proposal (Gunderson and Anderson, 1980) was to compute 95%-confidence limits on bias and TRSD separately and then compute the accuracy A at the confidence limits. Following Bonferroni (Miller, 1966), such an approach results in better than 90% confidence in the accuracy. However, the confidence level actually attained is generally excessively greater than 95% in this case of noncalibration.
2. The accuracy A surface may be approximated (ASTM, 1997) as a plane near the point of application if |bias| is large. Confidence limits on linear functions of bias and TRSD are easily computed (Johnson and Kotz, 1970) in terms of the noncentral Student-t distribution. Therefore, an approximate confidence limit on the accuracy is obtained. Equation (3) suggests a similar approximation (Appendix C).

MEAN SQUARE ERROR

Gaussian accuracy

This section unifies the concepts of symmetric-range accuracy A and Gauss' (1823) mean square error function, MSE. MSE (sometimes denoted as *standard measurement uncertainty*) is defined over $j = 1, \ldots, n$ representative measurements $\hat{c}_j$ of known C_j by:

$$\hat{\text{MSE}} \equiv \frac{1}{n}\sum_j (\hat{c}_j - C_j)^2 / C_j^2. \tag{9}$$

As is clear from its definition, MSE is an increasing function of both random imprecision in $\hat{c}_j$ and bias between $\hat{c}_j$ and C_j. Therefore, MSE provides another means of quantifying accuracy.

The explicit dependence of MSE on bias and imprecision can be found by computing the expected value of the estimate $\hat{\text{MSE}}$. If bias and imprecision TRSD are constant (as in the approximation above), then the expected value is:

$$\text{MSE} = \text{bias}^2 + \text{TRSD}^2. \tag{10}$$

Curves of constant MSE are shown (as circles) in Fig. 1.

Figure 1 demonstrates clearly that MSE and A are equivalent in the usual situation that $|\text{bias}| < \text{TRSD}/u_\alpha$. For example, Eqs (3) and (10) imply that at $\alpha = 95\%$,

$$\text{MSE} \approx (A/1.960)^2, \ |\text{bias}| < \text{TRSD}/1.645. \tag{11}$$

Knowledge of MSE gives A.

One advantage of MSE is its mathematical simplicity. For example, confidence limits on MSE can be easily estimated in the approximation that bias and imprecision TRSD are constant. For details, see Appendix C.

CONTINUAL SYSTEM EVALUATION

In some cases, the measurement system may be re-evaluated every time a measurement is taken. Pinning down the accuracy of each measurement is simplified in this case, at the expense of the required system evaluations. Simpler confidence intervals than tolerance intervals result.

As above, following the evaluation, the system may be calibrated for measurement of unknown concentration X. In this case, however, with continual re-evaluation, the following expression u is unit normal:

$$u = \frac{(\hat{X} - X)/X}{\text{TRSD} \times (C/c) \times \sqrt{1 + \frac{1}{n}}}. \tag{12}$$

With the intent of constructing a t-distributed variable via the ratio of a unit normal to a the square root of a chi-square variable divided by its number of degrees of freedom, the following expression is approximated as chi-square distributed:

$$\nu_{\text{eff}} \frac{\hat{\text{TRSD}}^2/\bar{c}^2}{\text{TRSD}^2/c^2} \approx \chi^2_{\nu_{\text{eff}}}. \tag{13}$$

The effective number of degrees of freedom ν_{eff} is determined by equating variances. The result is that, ignoring corrections of the order of TRSD^2/n,

$$\nu_{\text{eff}} \approx \nu. \tag{14}$$

Thus the following variable is approximately t-distributed:

$$t = \frac{(\hat{X} - X)/X}{\hat{\text{RSD}} \times \sqrt{1 + \frac{1}{n}}}. \tag{15}$$

Therefore, the following simple confidence limit results:

$$|\hat{X} - X|/X < t_{\frac{1}{2}(1+\alpha),\nu} \times \hat{\text{RSD}} \times \sqrt{1 + \frac{1}{n}} \tag{16}$$

at the α confidence level. For example, at $n = 16$ and $\alpha = 95\%$,

$$|\hat{X} - X|/X < 2.20 \times \hat{RSD}. \quad (17)$$

DISCUSSION

One of the themes of this paper has been the quantification of confidence in the results of evaluation experiments on given measurement systems. Two extremes as to how the evaluation is done were considered. On the one hand, the system undergoes one extensive evaluation prior to being applied many times. An advantage of this approach is that, given a carefully designed evaluation experiment, many of the contingencies possible in application of a method, perhaps in a hostile environment, may be covered. Sensitivity to environmental parameters (for example, ambient pressure) not measured during normal method application can be characterized. Furthermore, the effects can be minimized through calibration. The remaining uncertainty can be accounted for either as an environmentally related random error or as an extreme bias, so as to quantify how close to true values a measurement is expected to be.

Of course, practical realization of this scheme requires assurance that the measurement system in application behaves as when initially evaluated. In other words, the system must remain in a *state of statistical control* (Eisenhart, 1963). In order to accomplish this a quality assurance program is often implemented so as to monitor those aspects of the measurement process most likely to vary.

At the other extreme, a measurement system may be evaluated at each application. Such a scheme would likely find most use in measurements which can be taken under highly controlled conditions. An example would be the centralized preparation of calibration standards for use by diverse clients. As indicated in the paper, because of the extensive, continual evaluation effected, this approach results in confidence limits on what is to be measured with a *single* confidence level.

The two extremes as to method evaluation are therefore qualitatively different. With a single, albeit extensive, evaluation prior to multiple applications, *two* confidence levels are important. For any set of measurements, it can be stated that at specific confidence (for example, 95%) in the evaluation experiment, the range defined by the accuracy confidence limit ($A_{95\%}$) about the true measurand contains greater than a specific fraction (for example, 95%) of (future) measurements. Alternatively, $A_{95\%}$ can be used [see Eq. (4)] to define confidence limits about the true measurand so that at 95% confidence in the evaluation, the confidence limits contain the true value 95% of the time.

A separate theme has been the unification of supposedly different quantifications of accuracy. When method bias is under control, Gauss' mean square error, MSE, and the symmetric-range accuracy, A, have been found closely equivalent. This result simplifies calculation of both the accuracy A itself and its confidence limit. Furthermore, methods evaluated by different approaches are easily compared.

REFERENCES

ASTM. Standard practice for evaluating the performance of respirable aerosol samplers, D 6061. West Conshohocken: American Society for Testing and Materials; 1997.

ASTM. Standard practice for controlling and characterizing errors in weighing collected aerosols, D 6552. West Conshohocken: American Society for Testing and Materials; 2000.

Bartley DL. Inhalable aerosol samplers. Applied Occupational and Environmental Hygiene 1998;13:274–8.

Busch KA. SCP statistical protocol. In: Taylor DG, Kupel RE and Bryant JM, editors. Documentation of the NIOSH validation tests. Cincinnati: DHEW (NIOSH); 1977. [pub. no. 77-185].

Eisenhart C. Realistic evaluation of the precision and accuracy of instrument calibration systems. Journal of Research of the National Bureau of Standards — C. Engineering and Instrumentation 1963;76C(2).

Gauss CF. Theoria Combinationis Observationum Erroribus Minimis Obnoxiae Commentationes societatis regiae scientiarum Gottingensis recentiores, vol. V; 1823. p. 1–104. Gottingae [reprinted in Carl Friedrich Gauss Werke, band IV, Göttingen; 1873].

Gunderson EC, Anderson CC Development and validation of methods for sampling and analysis of workplace toxic substances., Springfield, VA: National Technical Information Service, 1980, ([DHEW (NIOSH) pub. no. 80-133, NTIS pub. PB-80-194-574]).

Hald A Statistical theory with engineering applications. New York: John Wiley & Sons, 1952.

ISO. Guide to the expression of uncertainty. Geneva: International Organization for Standardization, 1993.

ISO 16107. Workplace atmospheres — protocol for evaluating the performance of diffusive samplers. Geneva: International Organization for Standardization, 1999.

Johnson NL, Kotz S Continuous univariate distributions, vols 1 and 2. Boston: Houghton Mifflin Company, 1970, (ISBN 0471446262 (vol. 1), ISBN 0471584940 (vol. 2)).

Kennedy ER, Fischbach TJ, Song R, Eller PM, Shulman SA Guidelines for air sampling and analytical method development and evaluation., Cincinnati: DHHS (NIOSH), 1995, (pub. no. 95-117).

Kenny LC, Aitkin R, Chalmers C, Fabriès JF, Gonzales-Fernandez E, Kromhout H, Lidén G, Mark D, Riedinger G, Prodi V. A collaborative European study of personal inhalable aerosol sampler performance. Annals of Occupational Hygiene 1997;41:135–53.

Kenny LC, Bartley DL. The performance evaluation of aerosol samplers tested with monodisperse aerosols. Journal of Aerosol Science 1995;26(1):109–26.

Kenny LC, Lidén G. The application of performance standards to personal airborne dust samplers. Annals of Occupational Hygiene 1993;33:289–300.

Miller RG Jr. Simultaneous statistical inference., New York: McGraw-Hill, 1966.

NIOSH. Manual of analytical methods, Eller P, editor. 4th ed. US Department of Health and Human Services, 1994.

Wald A. Setting tolerance limits when the sample is large. Annals of Mathematical Statistics 1942;13:389–99.

Wald A. An extension of Wilk's method for setting tolerance limits. Annals of Mathematical Statistics 1943;14:45–55.

Wald A, Wolfowitz J. Tolerance limits for a normal distribution. Annals of Mathematical Statistics 1946;17:208–15.
Wilks SS. Determination of sample sizes for setting tolerance intervals. Annals of Mathematical Statistics 1941;12:91–6.
Wilks SS. Statistical prediction with special reference to the problem of tolerance intervals. Annals of Mathematical Statistics 1942;13:400–9.

APPENDIX A. ACCURACY FUNCTION

As defined in the body of this article, the accuracy A[bias, TRSD] at fixed bias and true relative standard deviation TRSD is the solution of:

$$\alpha = \Phi[(\text{bias} + A)/\text{TRSD}] - \Phi[(\text{bias} - A)/\text{TRSD}], \quad (A1)$$

where Φ denotes the cumulative normal function. The calculation of A is greatly simplified by utilizing the scaling properties of A, bias, and TRSD. Therefore, defining accuracy and bias ratios:

$$a \equiv A/\text{TRSD}, \ \delta \equiv \text{bias}/\text{TRSD}, \quad (A2)$$

Equation (A1) indicates that the accuracy ratio a is simply a function $a(\delta)$ (that is, of δ alone). The function $a(\delta)$ is most succinctly described by its differential equation, obtained by differentiating Eq. (A1):

$$\frac{da}{d\delta} = \tanh[a\delta]. \quad (A3)$$

Solutions of the limiting forms of Eq. (A3) are easily found with:

$$\tanh[a\delta] \rightarrow \left\{ \begin{matrix} a\delta, & a\delta \rightarrow 0 \\ \pm 1, & a\delta \rightarrow \pm\infty \end{matrix} \right\}. \quad (A4)$$

At $a\delta \rightarrow 0$,

$$a/a[0] \approx \exp\left[\tfrac{1}{2}\delta^2\right] \approx 1 + \tfrac{1}{2}\delta^2 \ (\delta \rightarrow 0), \quad (A5)$$

which translates to:

$$A \rightarrow u_{(1+\alpha)/2}\text{TRSD} \times \left[1 + \tfrac{1}{2}(\text{bias}/\text{TRSD})^2\right]. \quad (A6)$$

Therefore, ignoring factors of the order of $(\text{bias}/\text{TRSD})^4$,

$$A = u_{(1+\alpha)/2} \times [\text{bias}^2 + \text{TRSD}^2]^{1/2}, \quad (A7)$$

which corresponds to the central part of Eq. (3). The linear approximation in Eq. (3) is similarly obtained from the $a\delta \rightarrow \pm\infty$ limits of Eq. (A4).

APPENDIX B. EQUALITY OF ACCURACY CONFIDENCE LIMITS AND CLASSICAL TOLERANCE INTERVALS

That the interval $\frac{\bar{x}}{\bar{c}} \times [\bar{c} - ks, \bar{c} + ks]$ contains at least a fraction α of estimates $\hat{x}$ at probability γ means specifically that the integral I given by

$$I \equiv \frac{1}{\sqrt{2\pi} X\sigma/C} \int_{\frac{\bar{x}}{\bar{c}} \times (\bar{c} - ks)}^{\frac{\bar{x}}{\bar{c}} \times (\bar{c} + ks)} d\hat{x} \exp\left[-\frac{(\hat{x} - x)^2}{2X^2\sigma^2/C^2}\right] \quad (B1)$$

exceeds α at probability equal to γ.

Suppose the method is calibrated via the estimate of the reference concentration C, so that the calibrated estimates of X are given by $\hat{X}$:

$$\hat{X} = \frac{C}{\bar{c}}\hat{x}. \quad (B2)$$

Rewriting the above integral in terms of $\hat{X}$ then results in:

$$I = \frac{1}{\sqrt{2\pi} X\sigma/\bar{c}} \int_{X \times (1 - ks/\bar{c})}^{X \times (1 + ks/\bar{c})} d\hat{X} \exp\left[-\frac{\left(\hat{X} - \frac{C}{\bar{c}}x\right)^2}{2X^2\sigma^2/\bar{c}^2}\right]. \quad (B3)$$

As the integration range is symmetric about the true (unknown) concentration X, Eq. (1) gives the accuracy A_γ at confidence level γ directly by:

$$A_\gamma = k \times s/\bar{c} = k \times \hat{\text{RSD}} \approx \sqrt{\frac{\nu}{\chi^2_{1-\gamma,\nu}}} \times u_{(\alpha+1)/2} \times \left(1 + \tfrac{1}{2n}\right) \times \hat{\text{RSD}}, \quad (B4)$$

where $\hat{\text{RSD}}$ is the estimated relative standard deviation ($s/\bar{c}$).

APPENDIX C. CONFIDENCE LIMITS ON MSE

A confidence limit on MSE can be calculated by noting that $n \times \hat{\text{MSE}}/\text{TRSD}^2$ is distributed according to a well-researched probability density function, namely the noncentral chi-square distribution (Johnson and Kotz, 1970). In terms of the number of degrees of freedom n and the noncentrality parameter λ, the expected value and variance of the noncentral χ^2 are $n + \lambda$ and $2n + 4\lambda$, respectively. The parameter λ is given by:

$$\lambda = n \times \text{bias}^2/\text{TRSD}^2. \quad (C1)$$

As the noncentral chi-square distribution can be accurately approximated as proportional to the central chi-square distribution with an effective number of

degrees of freedom v_{eff} selected to give correct variances, the confidence limit MSE_γ on MSE is simple:

$$\text{MSE}_\gamma = \frac{v_{\text{eff}}}{\chi^2_{1-\gamma,v_{\text{eff}}}} \times \hat{\text{MSE}}, \tag{C2}$$

where v_{eff} is approximated by:

$$v_{\text{eff}} = n \frac{\hat{\text{MSE}}^2}{\hat{\text{TRSD}}^4 + 2\hat{\text{TRSD}}^2 \hat{\text{bias}}^2}. \tag{C3}$$

Note then that Eq. (3) gives a simple approximation to the accuracy confidence limit A_γ valid at arbitrary bias.

$$A_\gamma = \begin{cases} u_{(1+\alpha)/2} \times [\text{MSE}_\gamma]^{1/2}, & |\hat{\text{bias}}| < \hat{\text{TRSD}}/u_\alpha \\ |\hat{\text{bias}}| + t_{\gamma,v}[u_\alpha \times n^{1/2}] \times \hat{\text{TRSD}}/n^{1/2}, & \text{otherwise} \end{cases}. \tag{C4}$$

where $t_{\gamma,v}$ is the noncentral Student-t quantile. Numerical simulation indicates that Eq. (C4) results in between 95% and 96% at a target value equal to 95% at n=16, v=15.