SUMMARY OF THE
TNI ENVIRONMENTAL MEASUREMENT METHODS EXPERT COMMITTEE MEETING

OCTOBER 7, 2011

The Committee held a conference call on Thursday, October 7, 2011, at 2:00 pm EDT.

1 – Roll call

Richard Burrows, Test America (Lab)	Present
Brooke Connor, USGS (Other)	Absent
Dan Dickinson, NYSDOH (Accreditation Body)	Absent
Tim Fitzpatrick, Florida DEP (Lab)	Absent
Nancy Grams, Advanced Earth Technologists, Inc. (Other)	Present
Anand Mudambi, USEPA (Other)	Present
John Phillips, Ford Motor Co., (Other)	Present
Lee Wolf, Columbia Analytical Services (Lab)	Present
Ken Jackson, TNI administrative support staff	Absent

The following Associate Committee member was also present: Francoise Chauvin (NYC DOH), Bharat Chandramouli (Axys Analytical)

2 – Minutes from September 29, 2011

John Phillips made a motion to approve the minutes from the September 29 meeting which was seconded by Lee Wolf. All were in favor.

3 – Discussion – Incorporation of Proposed calibration language into existing TNI Quality Systems Standard

Prior to the call, Richard had circulated an amended draft standard dated October 7, 2011 (attached) and the TNI 2009 Quality Systems Standard Section 1.7 (Technical Requirements – Calibration).

Richard started the discussion by proposing that the committee consider modifying the calibration sections in the existing TNI Quality Systems Standard instead of creating a new calibration standard. The committee discussed this proposition and came up with the following pros and cons of this approach:

Pros: Advantages include the fact the users would need to go to only one place to implement/follow the standard (instead of two standard documents) and removing the need for separate standard maintenance.

Con: A potential disadvantage is that this would tie any calibration standard revisions to the Quality Systems revision schedule.

Nancy asked whether the committee would continue to develop the calibration guidance document. Richard said the guidance document is still needed to clarify items in the standard. He also stated

that any proposed changes to a TNI standard would have to be published 60 days before the January 2012 TNI meeting (sometime in November 2011).

The committee then held a formal vote on the following motion proposed by Nancy and seconded by Anand:

> *Motion: The items (material) on calibrations generated by the committee should be incorporated into the existing TNI Quality System Standard rather than creating a separate calibration standard.*
>
> In favor: Lee Wolf, Nancy Grams, Richard Burrows, John Phillips, Anand Mudambi (all the committee members that were present).
>
> Opposed: None

There also was strong agreement among the committee members that the guidance document should be completed first followed by extracting relevant sections for input into the Quality Systems standard.

4 – Discussion on Draft TNI Standard (dated October 7, 2011).

Richard reconfirmed the leads for the sections as follows:

Section 4.1: Brooke
Section 4.2: Lee
Sections 4.3, 4.3.1 and 4.3.2: Nancy
Sections 4.3.3, 4.3.4, 5.3.5, and 4.3.6: Tim
Sections: 4.3.7 and 4.4: John
Section 4.5: Richard
Section 4.6: Anand

Section 5.1, 5.2, and 5.2.1: Richard
Section 5.3: Anand
Sections 5.4 and 5.5; Brooke
Section 5.6: Tim
Sections 5.7 and 5.7.1: Tim
Section 5.7.2: Lee
Section 5.8: John
Section 6.0: Dan

Note: Lee suggested that Section 4.6 (Special Considerations for multi-response analytes) be refined after other sections are completed and Richard agreed.

Section 5.6 Discussion (Percent Relative Standard Error) Drafted by Tim

There was a lively discussion on this section starting with Nancy asking for clarification on the meaning of degrees of freedom in the Relative Standard Error (RSE) equation. Richard stated that this meant the number of calibration points. He further said that the discrete standards even at the

same calibration level are to be counted separately. Nancy felt that the points made by Richard need to be put in as clarifications (e.g., what each standard means in terms of the RSE equation, the fact that RSE cannot be used for single point calibrations, and that at least 3 points are needed for doing an RSE evaluation). Richard then pointed out that a zero point calibration cannot be counted as a standard in the equation. When asked by Nancy regarding running of blank standards during calibration, Richard stated that these could be run but could not be part of the RSE evaluation. Nancy will draft the clarifications for this section (including the following: if a zero point standard is run during an initial calibration sequence, it cannot be used in the RSE evaluation of the calibration curve. That means in the RSE equation in Section 5.6, n-p cannot be = 0 and x_i cannot be =0).

Section 5.7 Discussion (Residuals) Drafted by Tim

Richard said that there is a need for examples in this section to help users in diagnosing problems. John felt that the <50% criteria for the low standard was too wide. Richard stated that generally the low calibration standard is set at the quantitation limit. Nancy said that some clarification is needed regarding this point and whether there is a requirement to run a standard at the laboratory quantitation limit. John was also concerned about the spacing of standards and felt the criteria may be too wide if the laboratory was working in the high end of the calibration range. Richard said that it was rare to have such problems at the high end but there could be issues when some detectors get saturated. Nancy and John will draft a clarification after the low point calibration standard section is discussed.

Section 5.7.2 Discussion (Special Considerations for Multi Response Analytes) Drafted by Lee

Richard started the discussion by saying that generally for multi response analytes, each peak is separately calibrated (not aggregated), especially when there are interferences. Nancy wanted some recommendations on which method of calibration is better (separate vs. aggregate). John asked whether the calibration should be checked both ways. Francoise said that there should be language regarding meeting retention time windows for chromatographic methods. John pointed out that the calibration standard requires that standards meet retention time window and spectral id criteria. Nancy then brought up the use of peak areas vs. peak heights. Richard felt that this should be discussed in the calibration guidance document and requested that this discussion be placed in the parking lot (so it is not forgotten). Lee agreed to draft a clarification on the options for calibrating multi response analytes including a discussion on the pro and cons of separate vs. aggregate calibration depending on the type of samples expected.

Nancy then said it would be good to also discuss extracted standards (e.g., for volatile analyses) and said she would draft a short section on this topic.

5 – Next Steps

Anand will prepare the minutes for this meeting. The meeting adjourned at 3:30 pm. The next call is proposed for Friday November 4 at 11:00 am EDT.

LIST OF ACTION ITEMS

Item No.	Date Proposed	Action	Assigned to:	To be Completed by:
1	10/07/11	Clarifications on the use of RSE for calibration evaluation	Nancy	November 1, 2011
2	10/07/11	Criteria for low point standard.	Nancy and John	On hold till quantitation section is discussed
3	10/07/11	New Section on the use of peak areas vs. peak height: pros and cons	Lee	Placeholder
4	10/07/11	Multi analyte Responses: Discussion of calibration based on individual peaks or total peak area. Pros and cons based on types of samples expected.	Lee	November 1, 2011
5	10/07/11	Discussion of Extracted Standards	Nancy	November 1, 2011

Attachment

EL-V1M1-2011

ENVIRONMENTAL LABORATORY SECTOR

VOLUME 1

MANAGEMENT AND TECHNICAL REQUIREMENTS FOR LABORATORIES PERFORMING ENVIRONMENTAL ANALYSIS

Module x: Calibration

TNI Standard

P.O. Box 2439
Weatherford, TX 76086
817-598-1624
www.nelac-institute.org

This page intentionally left blank.

PREFACE

This Standard is the result of many hours of effort by those volunteers on The NELAC Institute (TNI) Proficiency Testing Committee. The TNI Board of Directors wishes to thank these committee members for their efforts in preparing this Standard as well as those TNI members who offered comments during the voting process.

This Standard may be used by any organization that wishes to implement a program for the accreditation of environmental laboratories.

VOLUME 1, MODULE 1

Calibration

1. INTRODUCTION, SCOPE AND APPLICABILITY

1.1. Introduction

xxx

1.2. Scope

xxx

1.3. Applicability

xxx

2. NORMATIVE REFERENCES

3. TERMS AND DEFINITIONS

For the purpose of this Standard, the relevant terms and definitions conform to *ISO/IEC 17011:2004* and *ISO/IEC 17025:2005*. Additional relevant terms are defined below.

4. CALIBRATION FOR ESTABLISHED METHODS

yy

4.1. External Standard Calibration (Brooke)

External standard calibration uses calibrants that are not added to the testing sample, but are contained in an individual and separate sample container for analysis. As such, they do not correct for individual sample losses, biases, or matrix effects. External standard calibrations are the most common form of calibration in the environmental laboratory.

Benefits of external standard calibration

The external standard calibration's popularity is due to its ease of use, its ubiquity in data reduction software, and its applicability to many methods. It works best when the sample preparation steps are limited so that variability of analytical results due to sample preparation is minimal; and it works well when sample injection into the analytical instrument is fairly reliable such that individual injection amounts do not vary significantly.

Drawbacks of external standard calibration

One drawback in using an external calibration is that it is separate from each sample, so it represents the instrument conditions at the time it was run, but not necessarily at the time when the samples are run. Instrument drift, loss of sensitivity, dirty samples, and matrix differences can all cause individual samples to exhibit increased variability that is unaccounted for with external calibration.

External calibration additionally assumes that the sample quantity has no error, which is not true. Further, external calibration assumes that the responses measured have random normally distributed error, which is only sometimes true.

Requirements when using external standard calibration

Comment [bfc1]: All of our sections should include this subsection, otherwise we aren't writing a standard, we are writing the guidance document.

External standard calibration must include three analyzed calibrants, including one at or near the reporting level or limit of quantitation (however prescribed in the methodology) and no higher than the linear range of the analysis. An exception is made for methods that specifically identify other calibration requirements. The linear range must be stated in the method, or determined and verified occasionally through the calibration verification procedures.

4.2. Internal Standard Calibration (Lee)

Internal standard calibration uses the relationship of detector responses from the target analyte and a specific standard of fixed concentration (internal standard). One or more internal standard is added to, or included in, each calibrant solution prior to analysis. A response factor (RF) for each target analyte is then established. In many cases the internal standards are specified or recommended in the method. Response factors are typically expressed as the ratio of the target compound response to the internal standard response, and taking into account concentrations. Certain methods (e.g. isotope dilution methods) may use unique equations for RF and the equation provided in the method must be used to determine the RF. The general equation for calculating the RF is as follows:

$$RF = (R_x C_{is})/(R_{is} C_x)$$

where: R_x = Response for analyte being measured.
A_{is} = Response for specific internal standard associated with the target analyte.
C_{is} = Concentration of the specific internal standard.
C_x = Concentration of the analyte being measured.

Using the same units for concentration, the RF is unitless. Response factors are determined for each analyte over the defined or established calibration range for the method. Using the RFs over the range, a suitable model is used to establish the calibration.

The same internal standard and concentration is added to each sample (or sample extract/digestate) prior to analysis, as described in the method. Quantitation of analytes in samples is then performed using the RF-based calibration and the responses from subsequent analyses.

4.2.1. Benefits of internal standard calibration

The primary benefit of using internal standard calibration is to improve the accuracy of the analysis by correcting for minor inconsistencies, physical or chromatographic interferences, or errors encountered during sample analysis.

To gain this benefit it is therefore important to match the analyte to a representative internal standard.

For many EPA methods, analysis of a “closing” continuing calibration verification standard is not necessary when using internal standard calibration.

4.2.2. Drawbacks of internal standard calibration

In internal standard calibration is not applicable to a wide range of analytical techniques. Internal standard calibration cannot be applied to wet chemistry or microbiology methods. The use of internal standard calibration is generally limited to chromatography, mass spectroscopy, and atomic emission spectroscopy.

It can be difficult to find an appropriate internal standard for certain analyses; that is, one that does not interfere with other measured components. This can be particularly true for chromatographic analyses by GC and HPLC without mass spectrometer detectors because of the inability to chromatographically resolve many internal standards from the target compounds.

When significant interferences are present in a sample suppression of the internal standard response can often be observed. An evaluation of this effect in sample analyses in included in data review or assessment.

4.2.3. Requirements when using internal standard calibration

When using internal standard calibration, any requirements for internal standards stated in the method must be used.

A method may specify or recommend the internal standards to be used. If the internal standards are not specified in the method, the analyst must select one or more internal standards that are similar in analytical behavior to the analytes of interest, and not expected to be found in samples. Related compounds or elements of similar analytical behavior and whose presence in environmental samples is highly unlikely must be used, such as brominated, fluorinated, stable isotopically labeled analogs; or for metals analyses elements not common to environmental samples. The internal standard and target analytes must not interfere with one another with the detection or determination of one another.

The internal standard approach can be used to ensure that adequate sensitivity exists for the analysis. Considering the internal standard concentration is kept equal for a given calibration, for chromatographic methods the RF can be assessed as a measure of sensitivity. The minimum RF criteria from the method must be used. Where no criteria exist, the minimum RF must be ≥ 0.05

4.3. Multipoint calibration

yy

4.3.1. Number of points

Related to range? Eg 3 per order of magnitude

Related to technique? Eg less for ICP than for GC?

Related to curve fit type

4.3.2. Spacing of points

Geometric?

4.3.3. Average Calibration or Response Factor (Tim)

The calibration factor is generally a term used for a certain type of external calibration procedure whereas the response factor is usually reserved when referring to a type of internal calibration procedure. Regardless of which type (external or internal) of calibration procedure is used, the concepts are similar and relatively straightforward. In the most general sense, the average calibration or response factor is merely the mean instrument response elicited per unit mass (or concentration) of analyte. The details of calculating an average response or calibration factor can be found in various texts or analytical methods (such as EPA Method 8000).

This type of calibration is the simplest model and should be considered first when establishing a method. A calibration model employing an average calibration or response factor assumes the calibration intercept of the curve is zero and that the instrument response is proportional to the analyte concentration (e.g., the slope of the response curve is constant). The use of average calibration or response factors is best employed for analytes and methods where there is no discernable instrument response for calibration blanks (i.e., chromatography methods) and where there is no curvature to the calibration model.

Average calibration or response factors should be used for quantification only when the relative standard deviation for individual standards is less than or equal to 20% (refer to Section 5.6). Ideally a minimum of five (5) calibration levels should be used calculate an average response or calibration factor and its precision, although some methods allow a minimum of three (3) levels.

4.3.4. Linear regression (Tim)

Linear regression is typically performed using a mathematical technique known as least squares. The technique involves fitting a line or curve to a set of data (typically 2-dimensional data having x and y coordinate values) in a way that minimizes the sum of the squares of the residuals – the difference between an actual datum point and that predicted by the resultant curve - while maximizing the likelihood of occurrence for the derived slope and intercept values (that is, in accordance with the maximum likelihood principle of statistics). The sum of the squares of the residuals is minimized when their partial derivatives with respect to the curve coefficients are equal to zero. This technique can be used to fit linear and non-linear curves to experimental data.

Linear regression assumes there are only two coefficients to the equation that describes the data set: a constant slope value and an intercept. The technique further assumes there are no errors in the concentrations of the calibrants; the only assumed error is in the instrument response to the calibrants. If a linear fit is employed, there should be *a priori* knowledge that the instrument response function with respect to concentration is indeed linear (or assumed to be linear) since the technique will derive a best-fit line regardless of the instrument response function. Linear regression can be used with and without weighting (discussed below). If weighting is not used, then the absolute error throughout the calibration range should be constant.

Linear regression is one of the most commonly used techniques for calibration. It's relatively simple, fits many experimental data sets and accounts for bias which can cause a non-zero intercept (i.e., where a calibration blank gives a non-zero instrument response). Simple un-weighted linear calibration models should not be used when the absolute uncertainty across the working range is not constant. Where accommodated by instrument software systems, weighted regression models should be used if errors are not constant (see Section 4.3.7).

4.3.5. Quadratic regression (Tim)

Not all methods (or detectors) exhibit a linear response with increasing concentration. To the extent possible, the working range of a method should be restricted to the area where response is linear with concentration. However, practical considerations or detector response functions may dictate that calibration data be fit to a non-linear curve. In such cases, a quadratic or second order polynomial curve may be employed. As with linear regression, the method of least squares is most often used to derive the best fit for calibration-response data by minimizing the square of the residuals in the response domain.

When a quadratic curve is used to model calibration data, the sensitivity of the response typically (but not always) decreases with concentration; essentially, the slope of the calibration curve approaches zero as concentration increases and the ability to discriminate concentration differences begins to degrade. The concentration at which the slope of the calibration curve is zero can be calculated by setting the first derivative of the derived calibration equation to zero and solving for concentration. Under no circumstances should any sample results be quantified in the region where the slope of the curve approaches zero. Whenever a quadratic calibration model is used, steps should be taken to ensure that there is sufficient resolving power throughout the working range to ensure that accuracy and precision targets for the method can be achieved.

4.3.6. Higher order regressions (Tim)

There are few cases where any calibration model greater than a second order curve is justified. Sometimes, however, there may be two or more phenomenon that, together cause a concentration-response model to represent a higher order polynomial. For example, reaction kinetic limitations at low concentrations and non-linear detector response may, in some cases, combine to yield a concentration-response model that is best approximated by a higher order polynomial. Once a signal is measured, solving third (and higher order)

calibration models for concentration may be complex, often requiring approximation techniques such as Newton's method, a Quotient-Difference algorithm or other numerical analysis techniques. Before resorting to the use of a calibration model having an order greater than two (quadratic), other, lesser order calibration models available within the instrument software system should be fully evaluated.

4.3.7. Weighting (John)

When variability changes continuously with concentration (non-constant standard deviation) weighting of the data may be beneficial. Weighting is particularly suitable if there is a systematic change in variability with respect to concentration, such as variability that increases with increasing concentration. A non-weighted calibration assumes a constant standard deviation, so that all data points influence the regression line equally (i.e. each point caries a "weight" of 1). However, if some data are noisier than others then the more variable points should not be allowed to have as much influence. This can be accomplished by evaluating the variability at each concentration and applying a weighting factor to each concentration point. The result is that the noisy responses have less influence on the calibration curve than do the precise values.

While computationally more complex, the failure to properly weight calibration data has two main effects on the quality of the calibration. First, the model's coefficient estimates for the slope and intercept will be noisy. Second, the prediction interval will be too wide in the well-behaved-data region and too narrow in the noisy-data region.

4.4. Single point calibration and blank calibration (John)

Some instruments and equipment respond linearly within the method operating range and single point calibration may be applicable. Examples include; some thermocouples, gages and photometric or spectroscopic equipment. It is important to remember that single point calibration is only applicable within the linear range of the method at concentrations where response bias is acceptable.

The advantage of single point calibration is that only a single calibration standard is required (the y-intercept is always assumed to be zero). The inherent simplicity of this technique makes it attractive; however extreme caution must be taken before concluding that a single point calibration is acceptable. It is still important to utilize an adequate number of calibration verification standards across the working range, since degradation in instrument performance is not often uniform across the operation range. Changes in instrument performance or response over time can affect both the slope and intercept of the calibration curve, which are assumed to be fixed with the single point calibration method

Since single point calibration depends upon the y-intercept being zero the trueness (lack of bias) of the calibration blank is critical. The measurement instrument must be properly zeroed (zero instrument response for zero sample concentration) to a true blank either physically, mechanically or electronically before proceeding with a single point calibration. It should be noted that the

further the lower end of the operating range is from zero the less critical the trueness of the calibration blank becomes.

4.5. Calibration for non-detects (Richard)

In most environmental analysis methods, particular multi-analyte tests, the usual result is a non-detect. This may mean that there is an instrument response, but the value determined is below the level chosen as a censoring limit by the laboratory, or it may mean that no instrument response at all was obtained. In either event, the actual quantitative value for a non-detect is meaningless and the value of the measurement is the ability to state that the analyte would have been detected (and reported) if it was present above a given level. For this reason, calibration acceptance criteria should be different for non-detected analytes than for detected analytes. Typical criteria useful for demonstrating the quantitative accuracy of detected analytes within the calibration range, such as RSD, RSE and % Drift have little value for non-detected analytes. Instead, a standard at the laboratory reporting limit that is detected when processed through the analytical procedure provides a level of confidence that the sensitivity of the method is sufficient to support the less than value reported by the laboratory.

4.6. Special considerations for multi-response analytes

Aroclors

Minimum five peaks which do not overlap (minimum resolution of 0.9) are needed for initial calibration. Additional consideration to select peaks that do not co-elute with interfering peaks or with organochlorine pesticide peaks.

Individual peak calibration vs summing responses into a single RF

For both approaches:

Each calibration peak should have a minimum resolution of 0.9 from other peaks chosen for calibration from the same multi response analyte. Minimum number of peaks needed for any analytes other than Aroclors?

Once the peak is identified, the peak is fully integrated across its intended retention time window.

Considerations:

The type of calibration option chosen will depends on the types of samples received, i.e.,. expected condition of the multi- component analyte. If little or no analyte degradation is expected, the individual peak calibration is preferable, whereas if analytes are expected to have significant degradation (e.g.., from weathered samples), the calibration response areas should be summed into a single RF.

5. INITIAL CALIBRATION ASSESSMENT

Comment [BR2]: Call this "Choosing the calibration type" or Building the calibration"?

yy

5.1. Rejection of calibration points (Richard)

Situations where rejection of points is/is not allowed

Reasons for rejection of points

5.2. Selection of the calibration curve type (Richard)

Most chromatographic software systems allow application of a wide selection of calibration curve types, including average response factor and linear and quadratic regressions. Various forms of weighting may be applied to linear and quadratic regressions. In order to determine the best curve fit to use for a particular calibration, the following factors should be considered:

The expected shape of the response curve and historical experience

A calibration that has historically been linear, and/or is expected to be linear that now exhibits a quadratic response may indicate the need for instrument maintenance

Method criteria

The selected curve must meet the calibration acceptance criteria in the method. However, meeting method criteria alone is commonly not sufficient to demonstrate that a particular calibration type is optimal.

Correlation coefficient (r) and Coefficient of Determination (r2)

These measures are frequently required to be evaluated for method compliance, but are not good indicators of the best curve to select. For example r and r2 will be higher for unweighted than for weighted curve fits to the same data set. However, it is quite likely that the unweighted curve will be a worse (for environmental analysis purposes) fit to the data due to large relative errors at the low end of the curve.

Relative standard Error (RSE)

The RSE evaluates the curve with equal weighting of relative residuals at each calibration level. The lowest RSE for a particular dataset indicates the calibration with the lowest overall relative error, making it a good measure for comparison of different curve fits

Individual residuals

The residuals at each point in the curve can be evaluated to compare calibration curve fits. This may be a significant task for methods with a large number of analytes, but is especially useful if it is known that the highest accuracy is desirable in a particular region of the calibration.

5.2.1. REQUIREMENTS FOR CURVE SELECTION Richard

5.2.1.1. Criteria in the method must be met

5.2.1.2. If unweighted regressions are used, the RSE or residuals of the low point in the calibration must be examined to ensure that unreasonable calibration error

is not present. The acceptable level for the RSE or residuals must be documented in the laboratory SOP.

5.3. Comparison to a separate source (Anand)

Definition of separate source

A standard obtained from a different manufacturer (not vendor) than the standard used for initial calibration. When only one manufacturer for the standard exists, a separate source standard is defined as being a different batch or lot than the standard used for initial calibration.

Which standards should be separate source

Once the initial calibration standards are run and evaluated, it is verified by running an initial calibration verification standard. Only the initial calibration verification standard has to be made from a second source standard.

5.4. Correlation coefficient and coefficient of determination (Brooke)

Two measures used for a measure of the degree of fit between instrument response and concentration are *r*, the correlation coefficient and r^2, the coefficient of determination. These two measures tell us slightly different things about the data, but more importantly don't tell us enough to conclude that our calibration regression will produce quality data. Generally, environmental analytical methods call for the use of *r* to assess calibrations. Unfortunately, *r* can be misleading, and its use can lead to poor results

Definition and uses of correlation coefficient (*r*)

The correlation coefficient (*r*) is a number that describes the goodness of fit between two variables. The correlation coefficient will result in a unitless number between -1 and +1, where +1 is a perfect positive fit of the two variables with a positive slope; zero is no correlation or a random relationship; and -1 is a perfect fit with a negative slope.

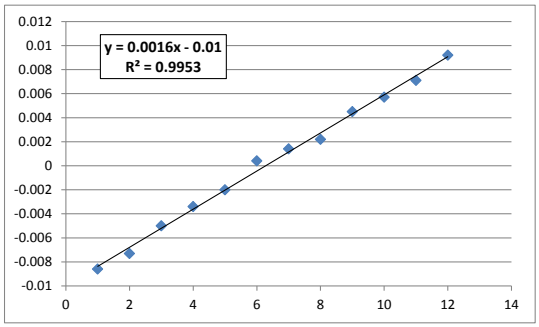
$$r = \frac{\sum xy - \sum x \sum y / n}{\sqrt{\sum x^2 - \left(\sum x\right)^2 / n}\sqrt{\sum y^2 - \left(\sum y\right)^2 / n}}$$

The use of *r* in the evaluation of a calibration is simple. The value obtained for *r* will give you an indication of the strength of a linear relationship and the direction of the relationship.

Comment [bfc3]: This is under "Requirements for Curve Selection". Don't we want to remove it from Requirements, and move it to Commonly Used Practices or something??

Comment [bfc4]: Why do some methods call for R2 and some call for R???

Comment [bfc5]: Needs variables defined!



5.4.1. Typical criteria for the correlation coefficient

Environmental analytical methods using the correlation coefficient to judge the quality of a calibration commonly use a correlation coefficient of 0.99 or greater as an acceptance limit.Benefits and limitations of r and r2

Benefits and limitations of *r*

The benefit of using the correlation coefficient is that it condenses the comparison of the instrument response versus concentration down to a single scalar, *r*. It is easy to use, easy to interpret, and it is used in most EPA analytical methods.

However, r does not imply causality. You can plot any two variables (such as the price of tea in China versus the number of bubbles in different pints of beer) and determine a correlation. It does not imply that one of the variables causes the results of the other.

The use of *r* as a measure of calibration acceptability has been widely accepted even though it has been challenged by many subject matter experts (Meier and Zund, 2000; Royal Society of Chemistry, Technical Brief; Taylor, 1990; Van Arendonk and Skogerboe, 1981). The problem with using *r* as an acceptance criterion for calibration is that it only indicates the variances from the averages. So you can still have a poorly fitted regression, with an acceptable *r* value

5.4.2. Definition and uses of coefficient of determination (r^2)

The coefficient of determination, r^2 , describes the percent of the variation that can be explained y the regression equation. It tells you how much variation in one variable is related to the variation in the other variable. A line that goes through every data point in a scatter plot is a line that accurately describes all the data. The r^2 in this case would be 1.00, meaning that 100 percent of the data are explained by the regression. An r^2 of zero means that you won't be able to predict a *y* value from a given *x* value at all. For example, a regression that has an *r* value of 0.902 means that 81% of the *y*-value data (because $0.902^2 = 0.81$) are explained by the regression but the remaining 19% are unexplained by the data.

$$r^2 = \left\{\left(\frac{1}{n}\right) * \sum\left[(x_i - \overline{x}) * (y_i - \overline{y})\right] / (\sigma_x * \sigma_y)\right\}^2$$

$N = number\ of observations$
$\sum = summation$
$x_i = the\ x\ value\ for\ observation\ i$
$\overline{x} = the\ mean\ x\ value$
$y_i = the\ y\ value\ for\ observation\ i$
$\overline{y} = the\ mean\ y\ value$
$\sigma_x = \ the\ standard\ deviation\ of\ x$
$\sigma_y = the\ standard\ deviation\ of\ y$

5.4.3. Typical criteria for the coefficient of determination (r^2)

The coefficient of determination typically is limited to 0.99 or greater for environmental analytical results (which is the same as an r value of 0.995).

5.4.4. Benefits and Limitations of (r2)

The benefit of the use of r^2 as a measure of goodness-of-fit is that it is easy to use, assess, and almost all analytical data reduction software packages include it in the default calculations. However, it is so easy to use that its correct interpretation is often ignored as long as the r^2 value exceeds the required minimum. This is especially true when the variability in the measure is due in major part to completely random events, where no amount of modeling can estimate its behavior. In this case, the r^2 value may indicate correlation of two events that are not related in the least. Its utility in providing a reliable measure of goodness-of-fit is not enough to consider it seriously as the sole measure.

5.5. Percent Relative Standard Deviation (Brooke)

5.5.1. Description of RSD

An evaluation of the relative standard deviation can be used to determine the suitability of using an average calibration response factor (for external calibration) or average relative response factor (for internal calibration) for quantifying sample signals. This calibration type is assumed to have a zero intercept and is equivalent to a weighted, least-square linear fit where the calibration is forced through zero and the weighting factor is 1/concentration^2.

The relative standard deviation is simply the square root of the variance of the calibration or response factors measured across the calibration range and expressed as a percentage of the average response. The variance is the sum of the squares of the 'residuals' – the difference between the observed calibration or response factors and the average factor – divided by the number of degrees of freedom. Mathematically, the RSD is determined as follows:

$$RSD = \frac{100}{\overline{CF}} \times \sqrt{\frac{\sum_{i=1}^{n}(CF_i - \overline{CF})^2}{n-1}}$$

or, for internal calibration

$$RSD = \frac{100}{\overline{RF}} \times \sqrt{\frac{\sum_{i=1}^{n}(RF_i - \overline{RF})^2}{n-1}}$$

where:

CFi ≡ the calibration response factor for individual calibrants,

$\overline{CF}$≡ the average or mean response factor,

RFi ≡ the relative response factor,

$\overline{RF}$ ≡ the average or mean response factor for individual calibrants,

n ≡ the number of calibration standards

5.5.2. Typical criteria

The precision of the average calibration or response factor derived from data collected throughout the working range is an important indicator of the suitability of the calibration model. If the precision is poor, the average calibration or response factor may not be the best choice to use for quantitation. Variability in response or calibration factors may be caused by:

- A non-zero intercept (presence of background), in which case a linear weighted calibration may be more appropriate
- A non-constant relationship between response and concentration, in which case a quadratic calibration may be more appropriate
- An inherent high degree of variability, in which case instrument maintenance or acceptance that the analyte is a relatively poor performer may be appropriate Determining what precision is acceptable varies across methods and programs. Some methods specify that the RSD of the calibration (or response) factors must be ≤ 10% (e.g., EPA Method 608), whereas others use ≤ 20% for acceptance criteria. Unless otherwise stated in a method or rule, when the calibration or response factor RSD exceeds 20%, the average calibration or response factor should (must?) not be used for quantitation unless other available calibration models are found to be less suitable.

5.5.3. REQUIREMENTS

5.5.3.1. In order for an average response factor calibration to be used, the RSE must be within the specifications in the analytical method

5.6. Percent Relative Standard Error (Tim)

5.6.1. Description of RSE

The relative standard error can be used to evaluate the fit of a variety of different calibration models (average responses, weighted and un-weighted first, second or n^{th} order models with and without a force through zero, etc.). The RSE is

similar to the RSD however this statistic utilizes the square root of the variance of the relative residuals - the difference between the measured response and the curve at each calibration point divided by the calibration level – expressed as a percentage. Therefore, equal weighting is given to residuals throughout the calibration range, making this statistic a robust tool for evaluating the applicability of calibration models. For the RSE, the number of degrees of freedom is adjusted for the number of variables in the calibration model chosen. Expressed mathematically, the RSE is defined as follows:Typical criteria

$$RSE = 100 \times \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{x_i' - x_i}{x_i}\right)^2}{n-p}}$$

x_i ≡ the true concentration of the standard at level i,
x_i' ≡ the predicted concentration at level i based on the calibration model chosen,
n ≡ the number of calibration points,
p ≡ the number of terms in the calibration model (e.g., 1 for a model using the average calibration or response factor or linear through origin, 2 for a first order linear fit not through the origin, 3 for a second order quadratic fit, etc.);

From the equation, it can be seen that the RSE requires a minimum number of calibration points, depending on curve type. For example, a second order (quadratic) calibration model would require at least 4 calibration points to compute the RSE. In this example, three calibration points is the minimum that can be used to define a quadratic fit, however no evaluation of fit can be performed.

5.6.2. REQUIREMENTS

5.6.2.1. **I**f the RSE is used to evaluate the calibration fit, the %RSE for the calibration selected must be within the method requirements (for RSE or RSD).

5.7. Residuals (Tim)

Residuals represent the difference between the measured response and that predicted by the calibration model. Normal distribution of residuals is important as is minimizing the residuals across the working range. Examination of the residuals associated with a calibration can be a good indicator of the adequacy of a given calibration model. Additionally, the assessment of residuals at the lower calibration level can be used to assess accuracy at lower quantitation levels.

A residual plot indicating a residual pattern is often used in statistical studies of calibrations, but is rather impractical for regular recurring analyses. Rather, the residuals at each point in the curve can be evaluated by quantitation of individual calibration points against the calibration curve.

Where method specifications do not exist, the recommended criteria are ≤50% difference for the low calibration point and ≤30% difference for all other points when each point is calculated back against the calibration.

5.7.1. Special considerations for the low point of the calibration

Recognizing that for many analyses the variation in response may become greater as lower levels of the analyte is measured, greater residuals can be expected. The actual residual (in terms of percent difference (%D) at the calibration point representing the quantitation limit) can be used as in indicator of the accuracy of the analysis at the quantitation limit. A lower %D will mean greater accuracy at low quantitation levels.

The recommended criteria are ≤50% difference for the low calibration point.

5.7.2. Special considerations for multi-response analytes Lee)

When performing calibrations for multi-response analytes (e.g. Aroclors) additional factors may impact the evaluation of residuals. For example, when a certain number of individual peaks are used to represent an analyte, each of these peaks will have differing responses, often significant differences. If using an approach where the representative individual peaks are calibrated one could not expect the same residual for each individual peak. When calibrating using the sum the individual responses into a single response factor representing the analyte, the residual is representative of the aggregate analyte rather than individual peaks or components.

5.8. With this consideration, and considering that the residual should indicate the adequacy of the calibration of the reported analyte, it is recommended that the residual evaluation for multi-response analytes be determined on the aggregate analyte basis as opposed to the individual peak or component basis. [Note: need to think this through some more and discuss (agreed). Are labs using summed responses to come up with one RF? Evaluation of single point calibrations (John)

5.8.1. Description and use of single point calibration

A single point calibration is actually a linear two point calibration with assumption that the calibration curve passes through zero. For accurate single point calibration results three things must be true; a) the intercept is truly zero, b) the measurement system is truly linear and c) the single point measurement has no error. Even if the single calibration point has no error (c) the further one gets away from the single point calibration concentration the greater the error of the result unless both (a) and (b) are also true. For this reason one must always have verification check standards at the calibration concentration as well as the lowest and highest operating concentrations. Expressed mathematically, the single point calibration equation is as follows:

$y = mx + b$

where,

y ≡ the measured concentration,

x ≡ the instrument response,

m ≡ the slope, i.e. calibration standard concentration divided by the instrument response,

b ≡ the intercept, assumed to be zero for single point calibrations.

Note: *It may be possible to use a single point calibration when the system has a non-linear response as long as the response can be transformed to a linear response. For example; a log rhythmic response can be transformed to a linear response by using the natural log of the response.*

5.8.2. Typical Criteria

The accuracy of a result at a given concentration within the operating range can be assessed by analyzing a verification standard and calculating the percent error at that concentration. For example; if the true concentration of the verification standard is 5.0 and the measured value based on the instrument response at that concentration is 4.5 then the percent error is the measured concentration minus the true concentration divided by the true concentration times one hundred (i.e., [(4.5-5.0)/5.0]*100 = -10%). In this example the result is biased low by 10%, since the percent error is a negative value. The percent error can not be calculated at zero concentration. The acceptance criteria for a single point calibration are dependent upon the measurement system and project MQOs, but 10% or less error across the method operating range is typical.

5.8.3. REQUIREMENTS

In order for a single point calibration to be used the percent error must be within the specifications in the analytical method across the measurement range.

6. CONTINUING CALIBRATION VERIFICATION

6.1. Frequency of the CCV (Dan)

6.2. Concentration of the CCV (Dan)

Discussion of benefits and problems associated with varying CCV concentration

6.3. Assessment of the CCV

Discussion of % difference and % drift

High failures vs. low failures

6.4. Special considerations for multi-response analytes

6.5. Special considerations for single point calibrations

7. SPECIAL TOPICS

7.1. Isotope dilution

7.2. Procedural standards