# Standard Practices for
# Infrared Multivariate Quantitative Analysis[1]

## 1. Scope

1.1 These practices cover a guide for the multivariate calibration of infrared spectrometers used in determining the physical or chemical characteristics of materials. These practices are applicable to analyses conducted in the near infrared (NIR) spectral region (roughly 780 to 2500 nm) through the mid infrared (MIR) spectral region (roughly 4000 to 400 $cm^{-1}$).

NOTE 1—While the practices described herein deal specifically with mid- and near-infrared analysis, much of the mathematical and procedural detail contained herein is also applicable for multivariate quantitative analysis done using other forms of spectroscopy. The user is cautioned that typical and best practices for multivariate quantitative analysis using other forms of spectroscopy may differ from practices described herein for mid- and near-infrared spectroscopies.

1.2 Procedures for collecting and treating data for developing IR calibrations are outlined. Definitions, terms, and calibration techniques are described. Criteria for validating the performance of the calibration model are described.

1.3 The implementation of these practices require that the IR spectrometer has been installed in compliance with the manufacturer's specifications. In addition, it assumes that, at the times of calibration and of validation, the analyzer is operating at the conditions specified by the manufacturer.

1.4 These practices cover techniques that are routinely applied in the near and mid infrared spectral regions for quantitative analysis. The practices outlined cover the general cases for coarse solids, fine ground solids, and liquids. All techniques covered require the use of a computer for data collection and analysis.

1.5 These practices provide a questionnaire against which multivariate calibrations can be examined to determine if they conform to the requirements defined herein.

1.6 For some multivariate spectroscopic analyses, interferences and matrix effects are sufficiently small that it is possible to calibrate using mixtures that contain substantially fewer chemical components than the samples that will ultimately be analyzed. While these surrogate methods generally make use of the multivariate mathematics described herein, they do not conform to procedures described herein, specifically with respect to the handling of outliers. Surrogate methods may indicate that they make use of the mathematics described herein, but they should not claim to follow the procedures described herein.

1.7 The values stated in SI units are to be regarded as standard. No other units of measurement are included in this standard.

1.8 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.*

## 2. Referenced Documents

2.1 *ASTM Standards:*[2]

D1265 Practice for Sampling Liquefied Petroleum (LP) Gases, Manual Method

D4057 Practice for Manual Sampling of Petroleum and Petroleum Products

D4177 Practice for Automatic Sampling of Petroleum and Petroleum Products

D4855 Practice for Comparing Test Methods (Withdrawn 2008)[3]

D6122 Practice for Validation of the Performance of Multivariate Online, At-Line, and Laboratory Infrared Spectrophotometer Based Analyzer Systems

D6299 Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance

D6300 Practice for Determination of Precision and Bias Data for Use in Test Methods for Petroleum Products and Lubricants

E131 Terminology Relating to Molecular Spectroscopy

---

---

## 3. Terminology

3.1 *Definitions*—For terminology related to molecular spectroscopic methods, refer to Terminology E131. For terminology relating to quality and statistics, refer to Terminology E456.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *analysis, n*—*in the context of this practice*, the process of applying the calibration model to a spectrum, preprocessed as required, so as to estimate a component concentration value or property.

3.2.2 *calibration, n*—a process used to create a model relating two types of measured data. In the context of this practice, a process for creating a model that relates component concentrations or properties to spectra for a set of known reference samples.

3.2.3 *calibration model, n*—the mathematical expression or the set of mathematical operations that relates component concentrations or properties to spectra for a set of reference samples.

3.2.4 *calibration samples, n*—the set of reference samples used for creating a calibration model. Reference component concentration or property values are known (measured by reference method) for the calibration samples and a calibration model is found which relates these values to the spectra during the calibration.

3.2.5 *estimate, n*—the value for a component concentration or property obtained by applying the calibration model for the analysis of an absorption spectrum.

3.2.6 *model validation, n*—the process of testing a calibration model with validation samples to determine bias between the estimates from the model and the reference method, and to test the agreement between estimates made with the model and the reference method.

3.2.7 *multivariate calibration, n*—a process for creating a model that relates component concentrations or properties to the absorbances of a set of known reference samples at more than one wavelength or frequency.

3.2.8 *reference method, n*—the analytical method that is used to estimate the reference component concentration or property value which is used in the calibration and validation procedures.

3.2.9 *reference values, n*—the component concentrations or property values for the calibration or validation samples which are measured by the reference analytical method.

3.2.10 *spectrometer/spectrophotometer qualification, n*—the procedures by which a user demonstrates that the performance of a specific spectrometer/spectrophotometer is adequate to conduct a multivariate analysis so as to obtain precision consistent with that specified in the method.

3.2.11 *surrogate calibration, n*—a multivariate calibration that is developed using a calibration set which consists of mixtures which contain substantially fewer chemical components than the samples which will ultimately be analyzed.

3.2.12 *surrogate method, n*—a standard test method that is based on a surrogate calibration.

3.2.13 *validation samples*—a set of samples used in validating the model. Validation samples are not part of the set of calibration samples. Reference component concentration or property values are known (measured by reference method), and are compared to those estimated using the model.

## 4. Summary of Practices

4.1 Multivariate mathematics is applied to correlate the spectra measured for a set of calibration samples to reference component concentrations or property values for the set of samples. The resultant multivariate calibration model is applied to the analysis of spectra of unknown samples to provide an estimate of the component concentration or property values for the unknown sample.

4.2 Multilinear regression (MLR), principal components regression (PCR), and partial least squares (PLS) are examples of multivariate mathematical techniques that are commonly used for the development of the calibration model. Other mathematical techniques are also used, but may not detect outliers, and may not be validated by the procedure described in these practices.

4.3 Statistical tests are applied to detect outliers during the development of the calibration model. Outliers include high leverage samples (samples whose spectra contribute a statistically significant fraction of one or more of the spectral variables used in the model), and samples whose reference values are inconsistent with the model.

4.4 Validation of the calibration model is performed by using the model to analyze a set of validation samples and statistically comparing the estimates for the validation samples to reference values measured for these samples, so as to test for bias in the model and for agreement of the model with the reference method.

4.5 Statistical tests are applied to detect when values estimated using the model represent extrapolation of the calibration.

4.6 Statistical expressions for calculating the repeatability of the infrared analysis and the expected agreement between the infrared analysis and the reference method are given.

## 5. Significance and Use

5.1 These practices can be used to establish the validity of the results obtained by an infrared (IR) spectrometer at the time the calibration is developed. The ongoing validation of estimates produced by analysis of unknown samples using the calibration model should be covered separately (see for example, Practice D6122).

5.2 These practices are intended for all users of infrared spectroscopy. Near-infrared spectroscopy is widely used for quantitative analysis. Many of the general principles described in these practices relate to the common modern practices of near-infrared spectroscopic analysis. While sampling methods and instrumentation may differ, the general calibration methodologies are equally applicable to mid-infrared spectroscopy. New techniques are under study that may enhance those discussed within these practices. Users will find these practices to be applicable to basic aspects of the technique, to include sample selection and preparation, instrument operation, and data interpretation.

5.3 The calibration procedures define the range over which measurements are valid and demonstrate whether or not the sensitivity and linearity of the analysis outputs are adequate for providing meaningful estimates of the specific physical or chemical characteristics of the types of materials for which the calibration is developed.

## 6. Overview of Multivariate Calibration

6.1 The practice of infrared multivariate quantitative analysis involves the following steps:

6.1.1 *Selecting the Calibration Set*—This set is also termed the training set or spectral library set. This set is to represent all of the chemical and physical variation normally encountered for routine analysis for the desired application. Selection of the calibration set is discussed in Section 17, after the statistical terms necessary to define the selection criteria have been defined.

6.1.2 *Determination of Concentrations or Properties, or Both, for Calibration Samples*—The chemical or physical properties, or both, of samples in the calibration set must be accurately and precisely measured by the reference method in order to accurately calibrate the infrared model for prediction of the unknown samples. Reference measurements are discussed in Section 9.

6.1.3 *The Collection of Infrared Spectra*—The collection of optical data must be performed with care so as to present calibration samples, validation samples, and prediction (unknown) samples for analysis in an alike manner. Variation in sample presentation technique among calibration, validation, and prediction samples will introduce variation and error which has not been modeled within the calibration. Infrared instrumentation is discussed in Section 7 and infrared spectral measurements in Section 8.

6.1.4 *Calculating the Mathematical Model*—The calculation of mathematical (calibration) models may involve a variety of data treatments and calibration algorithms. The more common linear techniques are discussed in Section 12. A variety of statistical techniques are used to evaluate and optimize the model. These techniques are described in Section 15. Statistics used to detect outliers in the calibration set are covered in Section 16.

6.1.5 *Validation of the Calibration Model*—Validation of the efficacy of a specific calibration model (equation) requires that the model be applied for the analysis of a separate set of test (validation) samples, and that the values predicted for these test samples be statistically compared to values obtained by the reference method. The statistical tests to be applied for validation of the model are discussed in Section 18.

6.1.6 *Application of the Model for the Analysis of Unknowns*—The mathematical model is applied to the spectra of unknown samples to estimate component concentrations or property values, or both, (see Section 13). Outlier statistics are used to detect when the analysis involves extrapolation of the model (see Section 16).

6.1.7 *Routine Analysis and Monitoring*—Once the efficacy of one or more calibration equations is established, the equations must be monitored for continued accuracy and precision. Simultaneously, the instrument performance must be monitored so as to trace any deterioration in performance to either the calibration model itself or to a failure in the instrumentation performance. Procedures for verifying the performance of the analysis are only outlined in Section 22. For petrochemicals, these procedures are covered in detail in Practice D6122. The use of Practice D6122 requires that a quality control procedure be established at the time the model is developed. The QC check sample is discussed in Section 22. For practices to compare reference methods and analyzer methods, refer to Practices D4855.

6.1.8 *Transfer of Calibrations*—Transferable calibrations are equations that can be transferred from the original instrument, where calibration data were collected, to other instruments where the calibrations are to be used to predict samples for routine analysis. In order for a calibration to be transferable it must perform prediction after transfer without a significant decrease in performance, as indicated by established statistical tests. In addition, statistical tests that are used to detect extrapolation of the model must be preserved during the transfer. Bias or slope adjustments, or both, are to be made after transfer only when statistically warranted. Calibration transfer, that is sometimes referred to as instrument standardization, is discussed in Section 21.

## 7. Infrared Instrumentation

7.1 A complete description of all applicable types of infrared instrumentation is beyond the scope of these practices. Only a general outline is given here.

7.2 The IR instrumentation is comprised of two categories, including instruments that acquire continuous spectral data over wavelength or frequency ranges (spectrophotometers), and those that only examine one or several discrete wavelengths or frequencies (photometers).

7.2.1 Photometers may have one or a series of wavelength filters and a single detector. These filters are mounted on a

turret wheel so that the individual wavelengths are presented to a single detector sequentially. Continuously variable filters may also be used in this fashion. These filters, either linear or circular, are moved past a slit to scan the wavelength being measured. Alternatively, photometers may have several monochromatic light sources, such as light-emitting diodes, that sequentially turn on and off.

7.3 Spectrophotometers can be classified, based upon the procedure by which light is separated into component wavelengths. Dispersive instruments generally use a diffraction grating to spatially disperse light into a continuum of wavelengths. In scanning-grating systems, the grating is rotated so that only a narrow band of wavelengths is transmitted to a single detector at any given time. Dispersion can occur before the sample (pre-dispersed) or after the sample (post-dispersed).

7.3.1 Spectrophotometers are also available where the wavelength selection is accomplished without moving parts, using a photodiode array detector. Post-dispersion is utilized. A grating can again provide this function, although other methods, such as a linear variable filter (LVF) accomplish the same purpose (a LVF is a multilayer filter that has variable thickness along its length, such that different wavelengths are transmitted at different positions). The photodiode array detector is used to acquire a continuous spectrum over wavelength without mechanical motion. The array detector is a compact aggregate of up to several thousand individual photodiode detectors. Each photodiode is located in a different spectral region of the dispersed light beam and detects a unique range of wavelengths.

7.3.2 The acousto-optical tunable filter is a continuous variant of the fixed filter photometer with no moving optical parts for wavelength selection. A birefrigent crystal (for example, tellurium oxide) is used, in which acoustic waves at a selected frequency are applied to select the wavelength band of light transmitted through the crystal. Variations in the acoustic frequency cause the crystal lattice spacing to change, that in turn, causes the crystal to act as a variable transmission diffraction grating for one wavelength (that is, a Bragg diffractor). A single detector is used to analyze the signal.

7.3.3 An additional category of spectrophotometers uses mathematical transformations to convert modulated light signals into spectral data. The most well-known example is the Fourier transform, that when applied to infrared (IR) is known as FT-IR. Light is divided into two beams whose relative paths are varied by use of a moving optical element (for example, either a moving mirror, or a moving wedge of a high refractive index material). The beams are recombined to produce an interference pattern that contains all of the wavelengths of interest. The interference pattern is mathematically converted into spectral data using the Fourier transform. The FT method can operate in the mid-IR and near-IR spectral regions. The FT instruments use a single detector.

7.3.4 A second type of transformation spectrophotometer uses the Hadamard transformation. Light is initially dispersed with a grating. Light then passes through a mask mounted on or adjacent to a single detector. The mask generates a series of patterns. For example, these patterns may be formed by electronically opening and shutting various locations, such as

in a liquid crystal display, or by moving an aperture or slit through the beam. These modulations alter the energy distribution incident upon the detector. A mathematical transformation is then used to convert the signal into spectral information.

7.4 Infrared instruments used in multivariate calibrations should be installed and operated in accordance with the instructions of the instrument manufacturer. Where applicable, the performance of the instrument should be tested at the time the calibration is conducted using procedures defined in the appropriate ASTM practice (see 2.1). The performance of the instrument should be monitored on a periodic basis using the same procedures. The monitoring procedure should detect changes in the performance of the instrument (relative to that seen during collection of the calibration spectra) that would affect the estimation made with the calibration model.

7.5 For most infrared quantitative applications involving complex matrices, it is a general consensus that scanning-type instruments (either dispersive or interferometer based) provide the greatest performance, due to the stability and reproducibility of modern instrumentation and to the greater amount of spectral data provided for computer interpretation. These data allow for greater calibration flexibility and additional options for selections of spectral areas less sensitive to band shifts and extraneous noise within the spectral signal. Scanning/interferometer-based systems also allow greater wavelength/frequency precision between instruments due to internal wavelength/frequency standardization techniques, and the possibilities of computer-generated spectral corrections. For example, scanning instruments have received approval for complex matrices, such as animal feed and forages (1, 2).[4]

7.6 Descriptions of instrumentation designs related to Refs (1) and (2) are found in Refs (3) and (4). Other instrumentation similar in performance to that described in these references is acceptable for all near-infrared techniques described in these practices.

7.7 For information describing the measurement of performance of ultraviolet, visible, and near infrared spectrophotometers, refer to Practice E275. For information describing the measurement of performance of dispersive infrared spectrophotometers, refer to Practice E932. For information describing the measurement performance of Fourier Transform mid-infrared spectrophotometers, refer to Practice E1421. For information describing the measurement performance of Fourier Transform near-infrared spectrophotometers, refer to Practice E1944. For spectrophotometers to which these practice do not apply, refer to Guide E1866.

## 8. Infrared Spectral Measurements

8.1 Multivariate calibrations are based on Beer's Law, namely, the absorbance of a homogeneous sample containing an absorbing substance is linearly proportional to the concentration of the absorbing species. The absorbance of a sample is defined as the logarithm to the base ten of the reciprocal of the transmittance, ($T$).

---

[4] The boldface numbers in parentheses refer to a list of references at the end of this standard.

$$A = \log_{10}(1/T)$$

The transmittance, $T$, is defined as the ratio of radiant power transmitted by the sample to the radiant power incident on the sample.

8.1.1 For measurements conducted by reflectance, the reflectance, $R$, is sometimes substituted for the transmittance $T$. The reflectance is defined as the ratio of the radiant power reflected by the sample to the radiant power incident on the sample.

Note 2—The relationship $A = \log_{10}(1/R)$ is not a definition, but rather an approximation designed to linearize the relationship between the measured reflectance, $R$, and the concentration of the absorbing species. For some applications, other linearization functions (for example, Kubelka-Munk) may be more appropriate (**5**).

8.1.2 For most types of instrumentation, the radiant power incident on the sample cannot be measured directly. Instead, a reference (background) measurement of the radiant power is made without the sample being present in the light beam.

Note 3—To avoid confusion, the reference measurement of the radiant power will be referred to as a background measurement, and the word reference will only be used to refer to measurements made by the reference method against which the infrared is to be calibrated. (See Section 9.)

8.1.3 A measurement is then conducted with the sample present, and the ratio, $T$, is calculated. The background measurement may be conducted in a variety of ways depending on the application and the instrumentation. The sample and its holder may be physically removed from the light beam and a background measurement made on the "empty beam". The sample holder (cell) may be emptied, and a background measurement may be taken through the "empty cell."

Note 4—For optically thin cells, care may be necessary to avoid optical interferences resulting from multiple internal reflections within the cell. For very thick cells, differences in the refractive index between the sample and the empty cell may change properties of the optical system, for example, shift focal points.

8.1.4 The sample holder (cell) may be filled with a liquid that has minimal absorption in the spectral range of interest, and the background measurement may be taken through the "background liquid." Alternatively, the light beam may be split or alternately passed through the sample and through an "empty beam," an "empty cell," or a "background liquid." For reflectance measurements, the reflectance of a material having minimal absorbance in the region of interest is generally used as the background measurement.

8.1.5 The particular background referencing scheme that is used may vary among instruments, and among applications. The same background referencing scheme must be employed for the measurement of all spectra of calibration samples, validation samples, and unknown samples to be analyzed.

8.2 Traditionally, a sample is manually brought to the instrument and placed in a cell or cuvette with windows that transmit in the region of interest. Alternatively, transfer pipes can be used to allow liquid to flow through an optical cell in the instrument for continuous analysis. With optical fibers, the sample can be analyzed remotely from the instrument. Radiation is sent to the sample through an optical fiber or bundle of

fibers and returned to the instrument by means of another fiber or bundle of fibers. Instruments have been developed that use single fibers to transmit and receive the radiation, as well as those using bundles of fibers for this purpose. Detectors and radiation sources external to the instrument can also be used, in which case only one fiber or bundle is needed. For spectral regions where transmitting fibers do not exist, the same function can be performed over limited distances using appropriate transfer optics.

Note 5—If the instrument uses predispersion of the light, some caution must be exercised to avoid introducing ambient light into the system at the sample position, since such light may be detected, giving rise to erroneous absorbance measurements.

8.3 Although most multivariate calibrations for liquids involve the direct measurement of transmitted light, alternative sampling technologies (for example, attenuated total reflectance) can also be employed. Transmittance measurements can be employed for some types of solids (for example, polymer films), whereas other solids (for example, powdered solids) are more commonly measured by diffuse reflectance techniques.

8.4 For most infrared instrumentation, a variety of adjustable parameters are available to control the collection and computation of the spectral data. These parameters control, for instance, the optical and digital resolution, and the rate of data acquisition (scan speed). A detailed description of the spectral acquisition parameters and their effect on multivariate calibrations is beyond the scope of these practices. However, it is essential that all adjustable parameters that control the collection and computation of spectral data be maintained constant for the collection of spectra of calibration samples, validation samples, and unknown samples for which estimates are to be made.

8.5 For definitions and further description of general infrared quantitative measurement techniques, refer to Practices E168. For a description of general techniques of infrared microanalysis, refer to Practice E334.

## 9. Reference Method and Reference Values

9.1 Infrared spectroscopy requires calibration to determine the proportionality relationship between the signals measured and the component concentrations or properties that are to be estimated. During the calibration, spectra are measured for samples for which these reference values are known, and the relationship between the sample absorbances and the reference values is determined. The proportionality relationship is then applied to the spectra of unknown samples to estimate the concentration or property values for the sample.

9.2 For simple mixtures containing only a few chemical components, it is generally possible to prepare mixtures that can serve as standards for the multivariate calibration of an infrared analysis. Because of potential interferences among the absorbances of the components, it is not sufficient to vary the concentration of only some of the mixture components, even when analyses for only one component are being developed. Instead, all components should be varied over a range representative of that expected for future unknown samples that are to be analyzed. Since infrared measurements are conducted on

a fixed volume of sample (for example, a fixed cell pathlength), it is preferable that concentration reference values be expressed in volumetric terms, for example, in volume percentage, grams per millilitre, moles per cubic centimetre, and so forth. Developing multivariate calibrations for reference concentrations expressed in other terms (for example, weight percentage) can lead to models that are linear approximations to what is really a nonlinear relationship and can lead to less accurate estimates of the concentrations.

9.3 For complex mixtures, such as those obtained from petrochemical processes, preparation of reference standards is generally impractical, and the multivariate calibration of an infrared analysis must typically be performed on actual process samples. In this case, the reference values used to calibrate the infrared analysis are obtained by a reference analytical method. The accuracy of a component concentration or property value estimated by a multivariate infrared analysis is highly dependent on the accuracy and precision of the reference values used in the calibration. The expected agreement between the infrared estimated values and those obtained from a single reference measurement can never exceed the repeatability of the reference method, since, even if the infrared estimated the true value, the measurement of agreement is limited by the precision of the reference values. Knowledge of the precision (repeatability) of the reference method is critical in the development of an infrared multivariate calibration. The precision of the reference data used in developing a model, and the accuracy of the model can be improved by averaging repeated reference measurements.

NOTE 6—If the reference values used to calibrate a multivariate infrared analysis are generated in a single laboratory, it is essential that the measurement process used to generate these values be monitored for bias and precision using suitable quality assurance procedures (see for example, Practice D6299. If primary standards are not available to allow the bias of the reference measurement process to be established, it is recommended that the laboratory participate in an interlaboratory cross-check program as a means of demonstrating accuracy.

NOTE 7—Samples like hydrocarbons from petrochemical process streams can degrade with time unless careful sampling and sample storage procedures are followed. It is critical that the composition of samples taken for laboratory or at-line infrared analysis, or for laboratory measurement of the reference data be representative of the process at the time the samples are taken, and that composition is maintained during storage and transport of the samples either to the analyzer or to the laboratory. Sampling should be done in accordance with methods like Practices D1265 and D4057, or Practice D4177, whichever are applicable. Whenever possible, sample storage for extended time periods is not recommended because of the likelihood of samples degrading with time in spite of sampling precautions taken. Degradation of samples can cause changes in the spectra measured by the analyzer and thus in the values estimated, and in the property or quality measured by the reference method.

9.4 If the reference method used to obtain reference values for the multivariate calibration is an established ASTM method, then repeatability and reproducibility data are included in the method. In this case, it is only necessary to demonstrate that the reference measurement is being practiced in accordance with the procedure described in the method, and that the repeatability obtained is statistically comparable to that published in the method. Data from established quality control procedures can be used to demonstrate that the repeatability of the reference method is within ASTM specifications. If such

data is not available, then repeatability data should be collected on at least three of the samples that are to be used in the calibration. These samples should be chosen to span the range of values over which the calibration is to be developed, one sample having a reference value in the bottom third of the range, one sample having a value in the middle third of the range, and one sample having a value in the upper third of the range. At least six reference measurements should be made on each sample. The standard deviation among the measurements should be calculated and compared to that expected based on the published repeatability.[5]

9.5 If the reference method to be used for the multivariate calibration is an established ASTM method, and the samples to be used in the calibration have been analyzed by a cooperative testing program (for example, octane values obtained from recognized exchange groups), then the reference values obtained by the cooperative testing program can be used directly, and the standard deviations established by the cooperative testing program can be used as the estimate of the precision of the reference data.

9.6 Reference methods that are not ASTM methods can be used for the multivariate calibration of infrared analyses, but in this case, it is the responsibility of the method developer to establish the precision of the reference method using procedures similar to those detailed in Practice E691, in the *Manual for Determining Precision for ASTM Methods on Petroleum Products and Lubricants*[5] and in Practice D6300.

9.7 When multiple reference measurements are made on an individual calibration or validation sample, a Dixon's Test (see A1.1) should be applied to the values to determine if all of the reference values came from the same population, or if one or more of the values is suspect and should be rejected.

## 10. Simple Procedure to Develop a Feasibility Calibration

10.1 For new applications, it is generally not known whether an adequate IR multivariate model can be developed. In this case, feasibility studies can be performed to determine if there is a relationship between the IR spectra and the component/property of interest, and whether a model of adequate precision could possibly be built. If the feasibility calibration is successful, then it can be expanded and validated. A feasibility calibration involves the following steps:

10.1.1 Approximately 30 to 50 samples are collected covering the entire range for the constituent/property of interest. Care should be exercised to avoid intercorrelations among major constituents unless such intercorrelations always exist in the materials being analyzed. The range in the concentration/property should be preferably five times, but not less than three times, the standard deviation of the reproducibility (reproducibility/2.77) of the reference analysis.

10.1.2 When collecting spectral data on these samples, variations in particle size, sample presentation, and process

conditions which are expected during analysis must be reproduced. Multiple spectra of the same sample under different conditions can be employed if such variations in conditions are anticipated during analysis.

10.1.3 Reference analyses on these samples are conducted using the accepted reference method. If the range for the component/property is not at least five times the standard deviation of the reproducibility for the reference analysis, then *r* replicate analyses should be conducted on each sample such that the $\sqrt{r}$ times the range is preferably five times, but at least three times, the standard deviation of the reference analysis.

10.1.4 A calibration model is developed using one or more of the mathematical techniques described in Sections 11 and 12. The calibration model is preferably tested using cross-validation methods such as SECV or PRESS (see 15.3.6). Other statistics can also be used to judge the overall quality of the calibration.

10.1.5 If the SECV value obtained from the cross validation suggests that a model of adequate precision can be built, then additional samples are collected to round out the calibration set, and to serve as a validation set, spectra of these samples are collected, a final model is developed, and validated as described in Sections 13, 14, and 15.

## 11. Data Preprocessing

11.1 Various types of data preprocessing algorithms can be applied to the spectral data prior to the development of a multivariate calibration model. For example, numerical derivatives of the spectra may be calculated using digital filtering algorithms to remove varying baselines. Such filtering generally causes a significant decrease in the spectral signal-to-noise. Digital filters may also be employed to smooth data, improving signal to noise at the expense of resolution. A complete description of all possible preprocessing methods is beyond the scope of these practices. For the purpose of these practices, preprocessing of the spectral data can be used if it produces a model which has acceptable precision and which passes the validation test described in Section 21. In addition, any spectral preprocessing method must be automated so as to provide an exactly reproducible result, and must be applied consistently to all calibration spectra, validation spectra, and to spectra of unknowns which are to be analyzed.

11.2 One type of preprocessing requires special mention. Mean-centering refers to a procedure in which the average of the calibration spectra (average absorption over the calibration spectra as a function of wavelength or frequency) is calculated and subtracted from the spectra of the individual calibration samples prior to the development of the model. The average reference value among the calibration samples is also calculated, and subtracted from the individual reference values for the calibration samples. The model is then built on the mean-centered data. If the spectral and reference value data are mean-centered prior to the development of the model, then:

11.2.1 When an unknown sample is analyzed, the average spectrum for the calibration site must be subtracted from the spectrum of the unknown prior to applying the mean-centered model, and the average reference value for the calibration set

must be added to the estimate from the mean-centered model to obtain the final estimate; and

11.2.2 The degrees of freedom used in calculating the standard error of calibration must be diminished by one to account for the degree of freedom used in calculating the average (see 15.2).

## 12. Multivariate Calibration Mathematics

12.1 Multivariate mathematical techniques are used to relate the spectra measured for a set of *calibration samples* to the reference values (property or component concentration values) obtained for this set of samples from a reference test. The object is to establish a multivariate *calibration model* that can be applied to the spectra of future, unknown, samples to estimate values (property or component concentration values). Only linear multivariate techniques are described in these practices; that is, it is assumed that the property or component concentration values can be modeled as a linear function of the sample spectra. Various nonlinear multivariate techniques have been developed, but have generally not been as widely used as the following linear techniques. These practices are not intended to compare or contrast among these techniques. For the purpose of these practices, the suitability of any specific mathematical technique should be judged only on the following two criteria:

12.1.1 The technique should be capable of producing a calibration model that can be validated as described in Section 18; and

12.1.2 The technique should be capable of providing statistics suitable for identifying if samples being analyzed are outside the range for which the model was developed; that is, when the estimated values represent extrapolation of the model (see 16.3).

NOTE 8—In the following derivations, matrices are indicated using boldface capital letters, vectors are indicated using boldface lowercase letters, and scalars are indicated using lowercase letters. Vectors are column vectors, and their transposes are row vectors. Italicized lowercase letters indicate matrix or vector dimensions.

12.1.3 All linear, multivariate techniques are designed to solve the same generic problem. If *n* calibration spectra are measured at *f* discrete wavelengths (or frequencies), then **X**, the spectral data matrix, is defined as an *f* by *n* matrix containing the spectra (or some function of the spectra produced by preprocessing, as described in Section 9) as columns. Similarly **y** is a vector of dimension *n* by 1 containing the reference values for the calibration samples. The object of the linear, multivariate modeling is to calculate a prediction vector **p** of dimension *f* by 1 that solves Eq 1:

$$\mathbf{y} = \mathbf{X^t p} + \mathbf{e} \qquad (1)$$

where $\mathbf{X^t}$ is the transpose of the matrix **X** obtained by interchanging the rows and columns of **X**. The error vector, **e**, is a vector of dimension *n* by 1, that is the difference between the reference values **y** and their estimates, $\hat{\mathbf{y}}$, where:

$$\hat{\mathbf{y}} = \mathbf{X^t p} \qquad (2)$$

12.1.4 For some applications, it may be useful to combine the spectral data with other measured variables (for example,

sample temperature, pH, mixing rates, etc.). These additional heterogeneous variable may simply be appended to the spectrum of each sample as if they were additional measured wavelengths. When heterogeneous data is used, it is important to consider the possibility that it may be appropriate to apply weighting factors to the heterogeneous variables in order to appropriately balance their influence on the calibration with respect to the influence of the spectral variables. Incorporation of additional heterogeneous variables in a model requires that these variables be measured for all future samples being analyzed using the model.

12.1.5 The estimation of the prediction vector **p** is generally calculated so as to minimize the sum of squares of the errors,

$$\mathbf{e^t e} = ||\,\mathbf{e^2}\,|| = (\mathbf{y} - \mathbf{X^t p})^t(\mathbf{y} - \mathbf{X^t p}) \tag{3}$$

Since **X** is generally not a square matrix, it cannot be directly inverted to solve Eq 3. Instead, the pseudo or generalized inverse of **X, X⁺**, is calculated as:

$$\mathbf{X^+ y} = (\mathbf{XX^t})^{-1}\mathbf{Xy} = \mathbf{p} \tag{4}$$

where **p** is the least square estimate of the prediction vector **p**. It should be noted that, in applying Eq 1-4, it is assumed that the errors in the spectral data in **X** are negligible compared to the errors in the reference data, and that there is a linear relationship between the component concentration or property and the spectral data. If either of these assumptions is incorrect, then the linear models derived here will not yield an optimal estimate of **p**.

12.1.6 In calculating the least square solution in Eq 4, it is assumed that the individual error values in **e** (see Eq 1) are normally distributed with common variance. This will be true if each of the individual reference values in **y** represents the result of a single reference measurement, and if the repeatability of the reference method is constant over the range of values in **y**. If the values in **y** represent averages of more than one reference method determination, then the least square expression in Eq 4 is not applicable. If $r_i$ reference values $y_{i1}, y_{i2}, y_{i3}$, ... $y_{ir}$ are measured for calibration sample $i$, then a weighted regression can be employed. If **R** is a diagonal matrix of dimension $n$ by $n$ containing the $r_i$ values for each of the calibration samples, then the weighted regression is given by:

$$\sqrt{\mathbf{R}}\ \bar{\mathbf{y}} = \sqrt{\mathbf{R}}\mathbf{X^t p} + \mathbf{e} \tag{5}$$

$$(\mathbf{XRX^t})^{-1}\mathbf{XR\bar{y}} = \mathbf{p} \tag{6}$$

where $\sqrt{\mathbf{R}}$ indicates the diagonal matrix containing the square roots of the $r_i$ values, and $\bar{\mathbf{y}}$ is the vector containing the averages of the $r_i$ reference values for each sample. If averages of multiple reference values are used in **y** and a weighted regression is used, special care must be taken to add back the variance removed by calculating the average reference values (see Section 11) so that the statistics for the model can be compared to those for a single reference value determination. The specific method in which the weighting is applied depends on the specific multivariate mathematics that are employed.

12.1.7 For most cases, if the calibration spectra are collected over an extended wavelength (or frequency) range, the number of individual absorption values per spectrum, $f$, will exceed the number of calibration spectra, $n$. In this case, the matrices $(\mathbf{XX^t})$ and $(\mathbf{XRX^t})$ are rank deficient and cannot be directly inverted. Even in cases where $f < n$, colinearity among the calibration spectra can cause $(\mathbf{XX^t})$ and $(\mathbf{XRX^t})$ to be nearly singular (to have a determinant that is near zero), and the direct use of Eq 4 and Eq 6 can produce an unstable model, that is, a model for which changes on the order of the spectral noise level produce significant changes in the estimated values. In order to solve Eq 4 and Eq 6, it is therefore necessary to reduce the dimensionality of **X** so that a stable inverse can be calculated. The various linear, mathematical techniques used for multivariate calibration are different means of reducing the dimensionality of **X** so as to be able to calculate stable inverses of $(\mathbf{XX^t})$ and $(\mathbf{XRX^t})$ and the estimate **p**.

12.2 *Multilinear Regression Analysis:*

12.2.1 In multilinear regression (MLR), a specific number, $k$, of individual wavelengths (or frequencies), or analytical regions, or both, are chosen such that $k <= n/6$. Depending on the particular application, the individual wavelength, or analytical regions may be individual wavelengths with or without baseline correction, or they may be a linear combination of several wavelengths (for example, ratios or integrated peak areas) with or without baseline correction. A new matrix **M** of dimension $k$ by $n$ is obtained from **X** by extracting the rows from **X** that correspond to the selected wavelengths (or frequencies). The calibration equation then becomes:

$$\mathbf{y} = \mathbf{M^t b} + \mathbf{e} \tag{7}$$

where **b** is a vector of dimension $k$ by 1 containing the set of regression coefficients defined at each of the chosen wavelengths (or frequencies). The solution for the regression coefficients is obtained as:

$$(\mathbf{MM^t})^{-1}\mathbf{My} = \mathbf{b} \tag{8}$$

If M comprises exclusively a subset of the wavelengths in X, the estimate of the full prediction vector, **p**, is obtained from **b** by substituting the values from **b** into the corresponding positions in **p** (corresponding to the selected wavelengths or frequencies), and setting all other elements of **p** (corresponding to the wavelengths or frequencies that were eliminated in going from **X** to **M**) to zero. If M comprises any integrated peak areas, then the value of $b$ which is substituted into the corresponding positions of **p** must be adjusted by dividing by the number of wavelengths incorporated into the integrated peak area. If baseline correction was applied to any wavelengths or integrated peak areas, then the same baseline correction must be applied to the spectrum measured for the unknown sample before applying the regression vector, **p**.

12.2.2 If a weighted regression is used, the corresponding form for Eq 8 becomes:

$$(\mathbf{MRM^t})^{-1}\mathbf{MRy} = \mathbf{b} \tag{9}$$

12.2.3 Not all commercial software packages that implement MLR include options for weighted regressions. If MLR models are developed with such packages, averages of multiple reference values should still be included in the **y** vector if they are available. The use of the average values will lead to better estimates of the regression coefficients, but the model produced will not be the least squares minimum. Standard errors of

calibration calculated by the software will generally not be meaningful in these cases since they are not expressed relative to a single reference measurement. Standard errors of calibration should be recalculated using the procedure described in Section 11.

12.2.4 The choice of the number of wavelengths (or frequencies), $k$, to use in multilinear regression is a critical factor in the model development. If too few wavelengths are used, a less precise model will be developed. If too many wavelengths are used, colinearity among the absorption values at these wavelengths may lead to an unstable model. The optimum number of wavelengths (or frequencies) for a model is related to the number of spectrally distinguishable components in the calibration spectra (see Section 15) and can generally only be determined by trial and error. As a rule, the number of wavelengths (or frequencies) used must be large enough to produce a model with the desired precision, but small enough to produce a stable model that passes validation.

12.2.5 The choice of specific wavelengths (or frequencies) to include in a multilinear regression model is also a critical factor in the model development. Several mathematical algorithms have been suggested for making this selection (6, 7, 8, 9). Alternatively, selection may be based on prior knowledge of a relationship between the absorptions measured and the property or component being modeled. It is beyond the scope of these practices to compare alternative selection methods. An adequate set of wavelengths (or frequencies) will, for the purpose of these practices, be defined as a set that produces a model with the desired precision that passes the validation test procedure described in Section 18.

12.3 *Principal Components Regression (PCR):*

12.3.1 Principal components regression (PCR) is based on the singular value decomposition of the spectral data matrix. The singular value decomposition takes the form:

$$\mathbf{X} = \mathbf{L}\sum\mathbf{S^t} \tag{10}$$

12.3.1.1 The scores matrix, $\mathbf{S}$, is a $n$ by $n$ matrix that satisfies the relationship:

$$\mathbf{S^tS} = \mathbf{I} \tag{11}$$

$$\mathbf{S^t(X^tX)S} = \Lambda \tag{12}$$

where $\mathbf{I}$ is a $n$ by $n$ identity matrix, and $\Lambda$ is the matrix of eigenvalues of $\mathbf{X^tX}$. The $n$ by $n$ matrix $\sum$ is the matrix of singular values, that are the square roots of the eigenvalues, that is:

$$\sum{}^2 = \Lambda \tag{13}$$

12.3.1.2 The loadings matrix, $\mathbf{L}$, is a $f$ by $n$ matrix that satisfies the relationships:

$$\mathbf{L^tL} = \mathbf{I} \tag{14}$$

$$\mathbf{L^t(XX^t)L} = \Lambda \tag{15}$$

12.3.1.3 The row vectors that make up the matrices $\mathbf{S}$ and $\mathbf{L}$ are orthonormal, that is, the dot product of the vector with itself is 1, and the dot product with any other vector in the matrix is 0.

NOTE 9—In some implementations of PCR, the data matrix $\mathbf{X}$ may be decomposed as the product of only two matrices, $\mathbf{S}$ and $\mathbf{L}$. Either $\mathbf{S}$ or $\mathbf{L}$

is then orthogonal but not orthonormal, and either $\mathbf{S^tS} = \Lambda$ or $\mathbf{L^tL} = \Lambda$.

12.3.1.4 Using the singular value decomposition, the pseudo inverse of the matrix $\mathbf{X}$ can be calculated as:

$$\mathbf{X^+} = \mathbf{S}\sum{}^{-1}\mathbf{L^t} \tag{16}$$

12.3.1.5 Using the pseudo inverse relationship in Eq 16, it is then possible to solve for the prediction vector $\mathbf{p}$. In practice, however, the full inverse of $\mathbf{X}$ as given in Eq 16 is not used, since it contains information relating to the spectral noise in the calibration spectra.

12.3.2 When a principal components analysis is conducted on a matrix $\mathbf{X}$ containing the calibration spectra, the signals arising from the calibration sample components generally account for the majority of the variance in $\mathbf{X}$, and are concentrated into the first $k$ loading vectors, that correspond to the larger eigenvalues. While the separation of signal and noise is seldom perfect, it is preferable to use only the first $k$ vectors in building a model. The singular value decomposition of $\mathbf{X}$ is then written as:

$$\mathbf{X} = \mathbf{L_a}\sum{}_{a}\mathbf{S_a^t} + \mathbf{L_n}\sum{}_{n}\mathbf{S_n^t} \tag{17}$$

where $\mathbf{S_a}$ is a $n$ by $k$ matrix containing the first $k$ columns of $\mathbf{S}$, $\mathbf{L_a}$ is a $f$ by $k$ matrix containing the first $k$ columns of $\mathbf{L}$, $\Sigma_a$ is a $k$ by $k$ diagonal matrix containing the first $k$ singular values, and $\mathbf{S_n}$, $\mathbf{L_n}$ and $\Sigma_n$ are the corresponding matrices containing the last $n-k$ elements of $\mathbf{S}$, $\mathbf{L}$ and $\Sigma$. The pseudo inverse of $\mathbf{X}$ is then approximated as:

$$\mathbf{X^+} = \mathbf{S_a}\sum{}_{a}^{-1}\mathbf{L_a^t} \tag{18}$$

12.3.2.1 The estimate for the prediction vector, $\mathbf{p}$, is then given as:

$$\mathbf{p} = \mathbf{L_a}\sum{}_{a}^{-1}\mathbf{S_a^t y} \tag{19}$$

12.3.2.2 Alternatively, the scores, $\mathbf{S}$, may be regressed against the reference values, $\mathbf{y}$, to obtain a set of regression coefficients, $\mathbf{b}$ :

$$\mathbf{y} = \mathbf{S_a b} + \mathbf{e} \tag{20}$$

$$\mathbf{b} = (\mathbf{S_a^t S_a})^{-1}\mathbf{S_a^t y} = \mathbf{S_a^t y} \tag{21}$$

12.3.2.3 Various stepwise regression algorithms (10, 11, 12) may be used to test which of the principal components (which columns in the scores matrix, $\mathbf{S}$) show a statistically significant correlation to the reference values in $\mathbf{y}$. Coefficients (elements of $\mathbf{b}$) for principal components that do not show a statistically significant correlation may be set to zero. The estimate for the prediction vector then becomes:

$$\mathbf{p} = \mathbf{L_a}\Sigma_a^{-1}\mathbf{b} \tag{22}$$

12.3.3 If the average of multiple reference measurements is used in the $\mathbf{y}$ vector, then a weighted regression should be used in calculating the prediction vector. The weighting is preferably applied to the scores in Eq 20 and Eq 21, and the spectra in $\mathbf{X}$ are not weighted prior to the singular value decomposition.

12.3.3.1 If $r_i$ individual reference values are measured for the $i^{th}$ calibration sample, then entering $r_i$ copies of the spectrum $x_i$ into the $\mathbf{X}$ matrix, or weighting the spectrum $x_i$ by $\sqrt{r_i}$ will alter the loadings that are calculated. If the spectrum $x_i$ is only measured once, the uncertainty in the spectral

variables contributed by $x_i$ is no different from that for the other $n - 1$ spectra. Weighting the spectrum $x_i$ prior to the singular value decomposition will tend to force noise characteristics of $x_i$ into the loadings, adversely affecting the model. Weighting the scores during the calculation of the regression coefficients will properly account for the differences in the variance among the components of the $\bar{\mathbf{y}}$ vector. The weighted regression equations become:

$$\sqrt{\mathbf{R}}\,\bar{\mathbf{y}} = \sqrt{\mathbf{R}}\mathbf{S_a}\mathbf{b} + \mathbf{e} \tag{23}$$

$$\mathbf{b} = (\mathbf{S_a^t}\,\mathbf{R}\mathbf{S_a})^{-1}\mathbf{R}\mathbf{S_a^t}\,\bar{\mathbf{y}} \tag{24}$$

12.3.4 Not all commercial software packages that implement PCR include options for weighted regressions. If PCR models are developed with such packages, averages of multiple reference values should still be included in the $\mathbf{y}$ vector if they are available. The use of the average values will lead to better estimates of the regression coefficients, but the model produced will not be the least squares minimum. Standard errors of calibration calculated by the software will generally not be meaningful in these cases since they are not expressed relative to a single reference measurement. Standard errors of calibration should be recalculated using the procedure described in 15.1.

12.3.5 As with wavelengths in multilinear regression, the choice of the number of principal components, $k$, to use in the regression is a critical factor in the model development. If too few principal components are used, a less precise model will be developed. If too many principal components are used, noise characteristics of the calibration samples will be incorporated into the model leading to unstable estimations. The optimum number of principal components for a model is related to the number of spectrally distinguishable components in the calibration spectra (see Section 15), and can generally only be determined by trial and error. As a rule, the number of principal components used must be large enough to produce a model with the desired precision, but small enough to produce a stable model that passes validation.

12.4 *Partial Least Squares (PLS):*

Note 10—The term PLS has been used to describe various mathematical algorithms. The version described here is a specific representation of the PLS-1 algorithm, and deals with only one set of reference values at a time. PLS-2 or multiblock PLS algorithms exist that can be used for the simultaneous calibration of multiple components or concentrations, or both. Except in special cases, PLS-2 generally produces calibrations which do not perform as well as those produced by PLS-1, and multiblock PLS involves complexities beyond the scope of these practices. Therefore, these practices do not address PLS-2 nor multiblock PLS algorithms. Various descriptions of the PLS-1 algorithm have been published (**13, 14, 15, 16, 17, 18, 19, 20**) many of which differ slightly in the actual computational steps. In implementing the PLS-1 algorithm, a choice must be made as to which, if either, of the scores or loadings vectors are to be normalized. In the following derivation, the scores vectors were normalized. If neither vector is normalized, or if the loadings vector is normalized instead of the scores vector, a different expression will be obtained for the prediction vector. Differences in the derivations should not result in significant differences in the numerical values obtained for the prediction vector, nor in estimates based on it.

12.4.1 Like PCR, PLS involves the decomposition of the spectral data matrix, $\mathbf{X}$, into the product of matrices. Unlike PCR where $\mathbf{X}$ is first decomposed, and then regressed versus the reference values, in PLS, the $\mathbf{y}$ vector is used in obtaining the decomposition of $\mathbf{X}$. The PLS proceeds by means of a series of steps, which are repeated in a loop. Each time the steps are repeated, a weighting vector $\mathbf{w_i}$ (of dimension $f$ by 1), a scores vector $\mathbf{s_i}$ (of dimension $n$ by 1), a regression coefficient $b_i$ (a scalar), and a loadings vector $\mathbf{l_i}$ (of dimension $f$ by 1) are calculated. The subscript $_i$ indicates the number of times the entire loop has been executed, and is initially 1.

12.4.1.1 *Step 1*—Calculation of a weighting vector of dimension $f$ by 1, $\mathbf{w_i}$:

$$\mathbf{X^t} = \mathbf{y}\mathbf{w_i^t} + \mathbf{Z} \tag{25}$$

$$\hat{w}_i = \mathbf{X}\mathbf{y} \tag{26}$$

12.4.1.2 *Step 2*—Scaling the weight vector $\hat{w}_i$ and calculation of a normalized scores vector, $s_i$, of dimension $n$ by 1:

$$\mathbf{X^t} = \mathbf{s}_i\hat{\mathbf{w}}_i^t + \mathbf{Z} \tag{27}$$

$$\hat{\mathbf{s}} = \mathbf{X^t}\,\hat{\mathbf{w}}_i \tag{28}$$

$$\hat{\mathbf{w}}_i = \hat{\mathbf{w}}_i(\hat{\mathbf{s}}^t\,\hat{\mathbf{s}})^{1/2} \tag{29}$$

$$\hat{\mathbf{s}}_i = \frac{\hat{\mathbf{s}}_i}{(\hat{\mathbf{s}}^t\,\hat{\mathbf{s}})^{1/2}} \tag{30}$$

12.4.1.3 *Step 3*—Regressing the scores vector against the reference values to obtain a regression coefficient, $b_i$:

$$\mathbf{y} = \hat{\mathbf{s}}_i b_i + \mathbf{e} \tag{31}$$

$$b_i = \hat{\mathbf{s}}_i^t\mathbf{y} \tag{32}$$

12.4.1.4 *Step 4*—Calculation of a loading vector, $\mathbf{l_i}$ of dimension $f$ by 1:

$$\mathbf{X} = \mathbf{l}_i\hat{\mathbf{s}}_i^t + \mathbf{Z} \tag{33}$$

$$\mathbf{l}_i = \mathbf{X}\hat{\mathbf{s}}_i \tag{34}$$

12.4.1.5 *Step 5*—Calculation of the residuals:

$$\mathbf{Z}_i = \mathbf{X} - \mathbf{l}_i\hat{\mathbf{s}}_i^t \tag{35}$$

$$\mathbf{e}_i = \mathbf{y} - b_i\hat{\mathbf{s}}_i \tag{36}$$

12.4.1.6 For subsequent times through the loop, the matrix $\mathbf{X}$ is replaced with the residuals matrix $\mathbf{Z}_{i-1}$ from the previous loop, and the $\mathbf{y}$ vector is replaced with the residuals vector $\mathbf{e}_{i-1}$. The loop is repeated $k$ times to obtain $k$ weighting, scores, and loading vectors, and $k$ regression coefficients. The overall expression for the results is then:

$$\mathbf{X} = \mathbf{L}\mathbf{S^t} + \mathbf{Z} \tag{37}$$

$$\mathbf{y} = \mathbf{S}\mathbf{b} + \mathbf{e} \tag{38}$$

where $\mathbf{S}$ is the $n$ by $k$ matrix containing the $\hat{\mathbf{s}}_i$ as rows, $\mathbf{L}$ is the $f$ by $k$ matrix containing the $\mathbf{l_i}$ as individual rows, $\mathbf{Z}$ is the residual from the spectral data matrix, and $\mathbf{e}$ is the residual from the estimation of the reference values. The estimate of the prediction vector is then given by:

$$\mathbf{p} = \mathbf{W}(\mathbf{L^t}\,\mathbf{W})^{-1}\mathbf{b} \tag{39}$$

12.4.2 If the values in the vector $\bar{\mathbf{y}}$ contain the average of multiple reference measurements, then a weighted regression should be employed in developing the model. Unfortunately, for PLS, development of an appropriate weighting scheme is complicated by the use of $\mathbf{y}$ in the decomposition of $\mathbf{X}$. If the spectrum $x_i$ corresponds to a sample for which $x_i$ reference

values are measured, then weighting both $\mathbf{X}$ and $\mathbf{y}$ by $\sqrt{\mathbf{R}}$ in Step 1 of the PLS algorithm will over emphasize the spectral variables contributed by $x_i$. Preferably, weighting is done only in the calculation of the regression coefficients in Step 3. Eq 31 and Eq 32 then become:

$$\sqrt{\mathbf{R}}\ \bar{\mathbf{y}} = \sqrt{\mathbf{R}}\hat{\mathbf{s}}_i b_i + \mathbf{e} \qquad (40)$$

$$b_i = (\hat{\mathbf{s}}_i^{\mathbf{t}}\mathbf{R}\hat{\mathbf{s}}_i)^{-1}\hat{\mathbf{s}}_i^{\mathbf{t}}R\ \bar{\mathbf{y}} \qquad (41)$$

12.4.2.1 The other steps in the algorithm proceed unchanged.

12.4.3 Not all commercial software packages that implement PLS include options for weighted regressions. If PLS models are developed with such packages, averages of multiple reference values should still be included in the $\bar{\mathbf{y}}$ vector if they are available. The use of the average values will lead to better estimates of the regression coefficients, but the model produced will not be the least squares minimum. Standard errors of calibration calculated by the software will generally not be meaningful in these cases since they are not expressed relative to a single reference measurement. Standard errors of calibration should be recalculated using the procedure described in 15.2.

12.5 *Frequency/Wavelength Selection in PCR/PLS Models:*

12.5.1 An important step in the development of PCR and PLS models is the selection of which frequencies/wavelengths to include in the model. The calibration model will not necessarily include data in the entire spectral range measured by the spectrometer. When analyzing specific analytes, spectral regions can be chosen based on knowledge as to where the analyte signals occur. When modeling physical or performance properties however, it may be advisable to include as much of the spectral range and information as practical. Spectral regions may be excluded from calibration models for a variety of reasons including high spectral noise, nonlinear spectrometer response and spectral interferences due to optical or environmental sources.

12.5.1.1 Spectrometers typically have limited range over which they will respond linearly. For dispersive spectrometers, stray light may limit the linear response range. Similarly, for FTIR spectrometers, phase errors can limit the linear response range. If spectral regions exhibiting nonlinear response are included in multivariate models, the number of variables needed to model the calibration data will increase. The nonlinearity in the X-Block may limit the transferability of the model between spectrometers (See Section 22), as well as the robustness of the model to spectrometer maintenance. Regions where samples are optically opaque should always be excluded from models.

12.5.1.2 Spectra can contain signals that are due to the spectrometer, sampling optics for the environment. For instance, for mid-IR, spectra often contain contributions from water vapor and carbon dioxide due to purge variations. The robustness of models can sometimes be improved by excluding spectral regions where such interferences occur. If regions with variable interference are included, it is important to either account for the variation or to take measures to remove the interference physically, for example, use a dry air purge to remove water vapor.

12.5.2 In selecting frequencies/wavelengths for inclusion in a model, it may be useful to calculate the average ($\bar{\mathbf{x}}$) and standard deviation ($\sigma_x$) of the calibration spectra. If $\mathbf{1}$ is a $n$ by 1 vector of ones, then:

$$\bar{\mathbf{x}} = \frac{\mathbf{X1}}{n} \qquad (42)$$

$$\sigma_{\mathbf{x}} = diag\left(\sqrt{\frac{(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^t)*(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^t)^t}{n-1}}\right) \qquad (43)$$

$r$ replicate spectra ($\mathbf{X}_r$) of one or more calibration samples should be collected, and similar average ($\bar{\mathbf{x}}_r$) and standard deviations ($\sigma_r$) values should be calculated.

$$\bar{\mathbf{x}}_r = \frac{\mathbf{X}_r \mathbf{1}}{r} \qquad (44)$$

$$\sigma_r = diag\left(\sqrt{\frac{(\mathbf{X}_r - \bar{\mathbf{x}}_r\mathbf{1}^t)*(\mathbf{X}_r - \bar{\mathbf{x}}_r\mathbf{1}^t)^t}{r-1}}\right) \qquad (45)$$

The ratio of $\sigma_r$ to $\sigma_x$ is useful in selecting frequencies/wavelengths for inclusion/exclusion from the model.

12.5.2.1 If, for a given frequency/wavelength, the ratio of $\sigma_r$ to $\sigma_x$ is less than 0.3, then the variation in the calibration set is significantly above the spectral noise level, and the frequency/wavelength should typically be included in the model.

12.5.2.2 If, for a given frequency/wavelength, the ratio of $\sigma_r$ to $\sigma_x$ is close to unity, then the variation in the calibration set is comparable to the spectral noise level, and the frequency/wavelength is a candidate for exclusion from the model.

12.5.3 During analyses, residuals are useful in detecting species that are not represented in the calibration samples. Exclusion of frequencies/wavelengths may reduce outlier detection capabilities of the model.

## 13. Estimation of Values from Spectra

13.1 If $\mathbf{x}$ (an $f$ by 1 vector) is the spectrum of a sample, then $\hat{y}$ (a scalar), the estimated component concentration or property value, is given by:

$$\hat{y} = \mathbf{x}^{\mathbf{t}}\mathbf{p} \qquad (46)$$

where $p$ is the prediction vector obtained from the multivariate calibration. The expression in Eq 46 involves only the dot product of two vectors to obtain the estimated value; it has the advantage of being computationally simple. However, alternative computations are often employed in obtaining $\hat{y}$, since they provide additional parameters required to calculate the uncertainty in the estimation as well as whether or not the estimation is being made by interpolation or extrapolation of the calibration model.

13.2 *Estimations by MLR*—For MLR, the values in $\mathbf{x}$ that correspond to the wavelengths (or frequencies) chosen in the calibration are extracted to form a vector $\mathbf{m}$ (of dimension $k$ by 1). The estimate $\hat{y}$ is then obtained as the dot product of the vector $\mathbf{m}$ with the vector of regression coefficients, $\mathbf{b}$:

$$\hat{y} = \mathbf{m}^{\mathbf{t}}\mathbf{b} \qquad (47)$$

13.3 *Estimations by PCR:*

13.3.1 For PCR, the vector $\mathbf{x}$ is first decomposed:

$$\mathbf{x}^{\mathbf{t}} = \mathbf{s}^{\mathbf{t}}\sum\mathbf{L}^{\mathbf{t}} \qquad (48)$$

$$\hat{\mathbf{s}}^{\mathbf{t}} = \mathbf{x}^{\mathbf{t}}\mathbf{L}\sum{}^{-1} \qquad (49)$$

13.3.2 The estimated scores, $\hat{\mathbf{s}}$ (a $k$ by 1 vector), are then multiplied by the regression coefficients obtained from the calibration to obtain $\hat{y}$:

$$\hat{\mathbf{y}} = \hat{\mathbf{s}}\mathbf{b} = \mathbf{x}^t\mathbf{L}\Sigma^{-1}\mathbf{b} = \mathbf{x}^t\mathbf{p} \qquad (50)$$

13.4 *Estimations by PLS:*

13.4.1 For PLS, the vector $\mathbf{x}$ is first decomposed in steps. Eq 51 and Eq 52 are repeated for each latent variable $i$ in the PLS model. After the first cycle through Eq 51 and Eq 52, $\mathbf{x}^t$ and $\mathbf{x}$ are replaced with $\mathbf{z}_{i\text{-}1}^t$ and $\mathbf{z}_{\text{-}i\text{-}1}$ from the previous cycle.

$$\hat{\mathbf{s}}_i^t = \mathbf{x}^t\mathbf{w}_i \qquad (51)$$

$$\mathbf{z}_i = \mathbf{x} - \mathbf{l}_i\hat{\mathbf{s}}_i^t \qquad (52)$$

13.4.2 The estimated scores, $\hat{\mathbf{s}}$ (a $k$ by 1 vector), are then multiplied by the regression coefficients obtained from the calibration to obtain $\hat{y}$

$$\hat{y} = \hat{\mathbf{s}}^t\mathbf{b} \qquad (53)$$

## 14. Post Processing

14.1 Several multivariate methods involve some post processing of the estimates from the multivariate model. The most common example is for mean-centered models (see Section 11), where the average reference value for the calibration set must be added to the initial estimate from the model to obtain the final estimate. A model can be developed to estimate changes in the pathlength of the cell used to contain the sample for analysis, and the estimated concentrations or property values can be scaled based on the results of the pathlength estimate.

14.2 A complete description of possible post-processing algorithms is beyond the scope of these practices. Post-processing can be employed if it provides a model with adequate precision, passes the validation test described in Section 18, and provided that the post-processing algorithm is automated, so as to provide exactly reproducible results, and is applied uniformly to the results from calibration, validation, and analyses.

## 15. Statistics Used in Evaluating and Optimizing Calibration Models

15.1 Various statistics are used to evaluate and optimize the performance of multivariate calibration models. These statistics are generally applied only to data in the calibration set; they should not be confused with the statistics that are used to validate the model (Section 18), that are calculated based on a separately analyzed validation set. The statistics discussed in this section are not appropriate for estimating the expected performance of a multivariate calibration model when used to estimate the values of unknown samples. The statistics discussed in Section 18 are solely appropriate for that purpose.

15.2 *Standard Error of Calibration:*

15.2.1 If $\hat{\mathbf{y}}$ are the values estimated for the calibration samples, and $\mathbf{y}$ are the corresponding reference values, then $\mathbf{e}$ is the calibration error vector defined by:

$$\mathbf{e} = \hat{\mathbf{y}} - \mathbf{y} \qquad (54)$$

The prediction errors include contributions from errors in the reference values for the calibration set, spectral errors in the spectra of the calibration set, and model errors (using wrong number of variables, nonlinear relationships, and so forth).

15.2.2 The standard error of calibration (SEC), is defined as:

$$SEC = \sqrt{\frac{\mathbf{e}^t\mathbf{e}}{d}} \qquad (55)$$

$d$ is the number of degrees of freedom in the calibration model. $d$ is typically equal to $n - k$, where $n$ is the number of calibration samples, and $k$ is the number of variables (wavelengths in MLR, principal components, or PLS latent variables) used in the model. If the spectral data and reference values are mean centered prior to the development of the calibration model (see Section 11), then $d = n - k - l$, since one degree of freedom is removed in calculating the averages. The standard error of calibration is sometimes referred to as the standard error of estimate (SEE).

NOTE 11—If a constant term is included in a MLR regression, or in the regression of PCR scores against concentrations or properties, then $d = n - k - l$, since one degree of freedom is associated with the constant. Care must be exercised in using a constant. In the case where neat samples are analyzed and the samples are run in fixed pathlength cells, the volume fractions of all components are constrained to sum to unity. Inclusion of a constant under these conditions can result in near singular matrices, and unstable models.

NOTE 12—For surrogate (see 17.6) calibrations, there is no relationship between the SEC calculated for the calibration based on the simple gravimetric mixtures and the error level expected when the model is used to analyze actual samples. It is recommended that such standard errors be subscripted as $SEC_{surrogate}$.

15.2.3 The standard error of calibration is used in estimating the expected agreement between values estimated using the calibration models and values that would be measured by the reference method (see Section 9). Some care must be applied in interpreting SEC if the values used in $\mathbf{y}$ are not single determinations by the reference method. If the values in $\bar{\mathbf{y}}$ for individual samples represent the average of multiple reference measurements, then the SEC calculated from Eq 55 is not on a per reference measurement basis. For example, if all values in $\bar{\mathbf{y}}$ are the average of three reference measurements, then the SEC calculated using Eq 55 can only be used to estimate the expected agreement between the infrared estimate and the average of three reference measurements.

15.2.3.1 If multiple reference values are used for some or all of the calibration samples, it is possible to calculate an SEC value that is on a per reference measurement basis. If $\mathbf{x}_i$ is the spectrum of the $i^{th}$ calibration sample, and $y_{i1}, y_{i2} \ldots y_{ir}$ are $r_i$ independently measured reference values for that sample, then the weighted regression Eq 9 for MLR, 23 and 24 for PCR, and 40 and 41 for PLS are preferably used in calculating the prediction vectors. Whether or not a weighted regression is employed, the variance removed by calculating the averages must be calculated as:

$$\sigma^2_{avg} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i)^2 \qquad (56)$$

where the first summation runs over all the $i = 1$ to $n$ calibration samples, the second summation runs over the $j = 1$ to $r_i$ reference values for the $i^{th}$ sample, and $\bar{y}_i$ is the average of the $r_i$ reference values for the $i^{th}$ sample. In this case, the standard error of calibration is calculated as:

$$SEC = \sqrt{\frac{\mathbf{e}^t\mathbf{Re} + \sigma^2_{avg}}{d_w}} \qquad (57)$$

15.2.3.2 The degrees of freedom for the weighted regression, $d_w$, are the total number of individual reference values measured for all the samples, minus the number of variables in the model:

$$d_w = \left[ \sum_{i=1}^{n} r_i \right] - k \tag{58}$$

15.2.3.3 If the spectral and reference data are mean centered prior to the development of the calibration, then:

$$d_w = \left[ \sum_{i=1}^{n} r_i \right] - k - 1 \tag{59}$$

The SEC calculated in this fashion will be on a per reference measurement basis.

15.2.3.4 An alternative expression for SEC in the case where multiple reference values per sample are used is given by:

$$\text{SEC} = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{r_i} (y_{ij} - \hat{y}_i)^2}{d_w}} \tag{60}$$

NOTE 13—In Eq 57, the **e** vector represents the difference between the estimated value and the reference value, where the reference value may be the average of more than one reference measurement. The matrix notation implies the sum of the weighted squares of the differences, where the square of the difference is weighted by the number of reference values that were included in the average. Alternatively, the square of the difference between the estimated value and each individual reference value can be computed and summed as in Eq 60, in which case the variance term is zero since the average reference values are not used in the calculation.

15.2.4 The standard error of calibration (SEC) is the standard deviation for the differences between reference and IR estimated values for samples within the calibration set. It is an indication of the total residual error due to the particular regression equation to which it applies. The SEC will generally decrease when the number of independent variables used in the model increases, indicating that increasing the number of terms will allow more variation in the data to be explained, or "fitted." The SEC statistic is a useful estimate of the theoretical "best" accuracy obtainable for a specified set of variables used to develop a calibration model.

15.3 *Optimizing the Number of Variables in a Model:*

15.3.1 Determining how many variables (wavelengths in MLR, principal components, or PLS latent variables) to use in a model is a critical step in the model development. Unfortunately, there are no hard and fast rules upon which to make this determination. In general, if too few variables are used, a less precise model will result. If too many variables are used, the estimates from the model may be unstable, that is, small changes in the spectrum on the order of the spectral noise may produce statistically significant changes in the estimates.

15.3.2 The maximum number of variables that should be used in developing a multivariate calibration model, $k$, is related to the number of detectable, spectrally distinguishable components (or functionalities) that are present in the calibration set. Components (or functionalities) are spectrally distinguishable if they give rise to absorptions which are not linearly correlated among the calibration samples, and if the change in the absorptions among the calibration spectra is larger than the spectral noise. If, within a calibration set, the concentrations of

components are linearly correlated, then the absorptions due to these components will also be linearly correlated. Even if these components have isolated absorption features, they will not be spectrally distinguishable to the multivariate mathematics, and will contribute at most one variable to the multivariate model. If the concentrations of the components are nearly correlated, such that the absorptions due to the components are colinear to within the spectral noise, then the components are not spectrally distinguishable. If components are present at sufficiently low levels so that the component absorption is below the spectral noise, then the component is not spectrally detectable and cannot contribute a variable to the multivariate model. Clearly, for complex mixtures, the number of detectable, spectrally distinguishable components (or functionalities) is often less than the number of real chemical components. There are, however, other sources of spectral variation that are not linearly correlated to the concentrations sample components but which will act as a spectrally distinguishable component. Some examples of other potential sources of spectral variation are changes in pH, particle size, sample temperature, sample placement, sampling procedure, chemical interactions among the sample components, spectral non-linearities or band shifts, instrument power supply voltages, ambient temperature, time of day, month, season, or year. Note that this list is not exhaustive.

15.3.3 One method of estimating the maximum number of detectable, spectrally distinguishable components among a set of calibration spectra requires knowledge of the spectral noise level. The spectral noise level can be estimated from replicate measurements conducted on a single sample. For instance, if replicate spectra are conducted on one sample, a PCR analysis of the spectra can be conducted. Since the spectra all represent the same material, only one principal component should be present in the spectral data. The percentage of the variance due to the first principal component (the first eigenvalue divided by the sum of all the eigenvalues) can be calculated. This percentage of the variance can be used to estimate a cutoff point for determining how many principal components to include in a model, namely, the sum of the first $k$ eigenvalues divided by the sum of all the eigenvalues should be of the same order as the cutoff. Similar calculations can be performed using PLS. For MLR, tests for colinearity among the absorbances at candidate wavelengths are generally conducted as part of the wavelength selection procedure. For instance, if a model is built using $k$ wavelengths for which the signals are linearly independent, the linear dependence of all candidate wavelengths for inclusion in a model based on $k + 1$ wavelengths can be checked. If the signals at all candidate wavelengths can be fit as a linear combination of the $k$ wavelengths already selected to within the spectral noise level, then $k$ is generally the maximum number of linearly independent wavelengths upon which an optimum model would be based.

15.3.4 Models can be built using other than $k$ variables, provided that such models exhibit adequate precision and pass validation.

15.3.5 Knowledge of the precision of the reference method is also useful in determining how many variables to include in a multivariate model. As discussed, the agreement between

infrared estimated values and reference values can never exceed the repeatability of the reference method, since, even if the infrared estimated the true value, the measure of the agreement would be limited by the repeatability of the reference method. Comparison of the standard error of calibration (calculated on the basis of a single reference measurement) against the standard deviation calculated from the reference method repeatability provides an indication of the maximum number of variables to include in a model. Standard errors of calibration that are lower than the standard deviation for the reference method indicate overfitting of the data.

15.3.6 Cross validation procedures are also used to estimate the optimum number of variables that should be included in a model. In cross validation, one or more sample spectra are removed from the data matrix, their corresponding reference values are removed from the reference value vector, and a model is built on the remaining samples. The model is then used to estimate the value for the samples that were left out. This process is repeated until each sample has been left out once. The error from the cross validation, $\mathbf{e_{cv}}$, is then calculated as

$$\mathbf{e_{cv}} = \hat{\mathbf{y}}_{cv} - \mathbf{y} \qquad (61)$$

where $\hat{\mathbf{y}}_{cv}$ is the vector containing the cross validation estimates. A PRESS value can then be calculated as:

$$\text{PRESS} = \mathbf{e_{cv}^t}\mathbf{e_{cv}} = \sum_{i=1}^{n} \left( \hat{y}_{cv_i} - y_i \right)^2 \qquad (62)$$

15.3.6.1 A standard error of cross validation (SECV) is calculated as:

$$\text{SECV} = \sqrt{\frac{\text{PRESS}}{n}} \qquad (63)$$

15.3.6.2 PRESS or SECV values can be calculated as a function of the number of variables used in the model. The procedure would normally start by using one variable as a model while leaving a single sample out of the calibration set. After a calibration is developed using the remaining samples, the algorithm predicts the excluded sample and records the difference between the reference and estimated values. This procedure is iterated (repeated) for the entire sample set, and the PRESS (or SECV) value for the one variable model is reported. The procedure then adds another variable and repeats the process. The PRESS procedure will stop when the predesignated number of factors is reached (say 10 to 20 maximum). The calibration model with the smallest PRESS (SECV) can be selected as the optimum model for the calibration set used. If more than one model have similar PRESS values, the one with fewer variables will generally be chosen.

15.3.6.3 A plot of PRESS (SECV) values (*y*-axis) versus the number of variables (*x*-axis) is often used to determine the minimum PRESS corresponding with the optimum number of variables in the calibration model. A minimum in the function can be taken as an indication of the maximum number of variables to be used. If no minimum occurs, the first point at which the PRESS or SECV reaches a more or less constant level can provide an indication of the maximum number of variables to include. Comparisons of SECV against the standard deviation for the reference method repeatability are again

useful, SECVs significantly lower than the standard deviation suggesting overfitting of the data.

15.3.6.4 An excellent description of the cross validation procedure (algorithm) is found in page 325 of Ref (**21**). Calculation of PRESS and SECV can be computationally intensive and can result in the use of substantial computer time.

NOTE 14—The exact values of PRESS and SECV calculated will depend on how many samples are left out during each cycle of the cross validation. If more than one sample is left out during a cycle, then the PRESS and SECV will depend on the combination of samples left out. Cross validation routines that leave out multiple spectra during each cycle require less computation time than routines that leave out one spectrum at a time. However, the results of such routines are less comparable and reproducible than those which leave out one spectrum at a time.

15.3.7 The above-mentioned methods for estimating the number of variables to use in a model are intended only as guidelines. None of the methods can be relied upon to always produce a stable model. The ultimate test for the number of variables is whether or not the model can be validated as described below. The number of variables used in a model must ultimately be chosen to produce a model with the desired precision that can be validated.

15.4 *Confidence Limits for an Estimated Value:*

15.4.1 The confidence limits for a value estimated by a multivariate model is given by:

$$t \cdot \text{SEC} \cdot \sqrt{1+h} \qquad (64)$$

where t is the student's t value for the number of degrees of freedom in the model, and *h* is the leverage statistic defined in 16.2. If t values are chosen from Table A1.3 for the 95 % probability level, then for a validated model, a single value measured by the reference method is expected to fall within a range from of $\hat{y} - t \cdot \text{SEC} \cdot \sqrt{1+h}$ to $\hat{y} + t \cdot \text{SEC} \cdot f\sqrt{1+h}$ for 95 % of samples analyzed, provided that the analysis is an interpolation of the model. The confidence limits for an estimated value in Eq 64 are sometimes referred to as the confidence bands or confidence intervals for the estimate.

15.4.2 The use of Eq 64 to estimate the confidence limits is only an approximation since it ignores any uncertainty in **x**, the spectral data. The confidence limits in Eq 64 derive from the assumption that the errors in **x** are negligible compared to the errors in **y**, and that the spectrum **x** can be completely described by the variables used in the model. If the errors in the spectral data are not negligible, or if the spectrum **x** contains absorptions due to components that were not present in the calibration set, the confidence limits in Eq 64 underestimate the potential error in the estimate. Eq 64 is expected to give a reasonable approximation for the confidence limits on an estimated value for samples that are interpolations of the model (see 16.4).

15.5 *Additional Statistics for Evaluating the Mathematical Models:*

15.5.1 A variety of statistical tests are in use for evaluating calibration models. Some tests that are in common use include:

15.5.1.1 Coefficient of multiple determination,

15.5.1.2 Correlation coefficient,

15.5.1.3 F-test statistic (F for regression),

15.5.1.4 Partial F or $t^2$ test for a regression coefficient,

15.5.1.5 Standard error of calibration (standard error of estimate), and

15.5.1.6 Bias corrected standard error.

15.5.2 Although many of these tests have been more commonly applied to MLR models, some are equally applicable to PCR and PLS models. Details on these tests and related statistical terms are included in Annex A2. Further explanations for these statistical tests can be found in Annex A2 and several references (**22, 23**).

## 16. Outlier Statistics

16.1 During calibration, outlier statistics are applied to identify samples that have unusually high leverage on the multivariate regression. During analysis, outlier statistics are employed to detect samples which represent an extrapolation of the model.

16.2 *Leverage Statistic*—The leverage statistic, $h$, is a scalar measure of where the spectral vector **x** lies within the multivariate parameter space used in the model. The leverage statistic is used in detecting outliers during the calibration, in detecting extrapolation of the model during analyses, and in estimating the uncertainty on an estimated value.

NOTE 15—Commercial software packages use numerous variations on the leverage statistic. The leverage statistic is sometimes referred to as the hat matrix (**24**), or as the Mahalanobis Distance, $D^2$ (although it is actually the square of the distance). Various commercial software packages may use $D$ instead of $D^2$. Some software packages may scale $h$ (or $D^2$) by $n$ (or $n-1$ if mean-centered), to obtain a statistic that is independent of the number of calibration samples. If this scaled statistic is further multiplied by $(n-k-1)/nk$, a statistic that has an $F$ distribution is obtained (**25**). The leverage statistic, $h$, is preferred here since it is easily related to the number of samples and variables. Model developers should attempt to verify exactly what is being calculated. Both mean-centered and not mean centered definitions for $h$ exist, with the mean-centered approach preferred. Regardless of whether mean centering of data is performed, the statistic designated $h$ has valid utility for outlier detection.

16.2.1 If **x** is a spectral vector (dimension $f$ by 1) and **X** is the matrix of calibration spectra (of dimension $f \times n$), then the leverage statistic is defined as:

$$h = \mathbf{x^t(XX^t)^+x} \qquad (65)$$

16.2.2 For a mean-centered calibration, **x** and **X** in Eq 65 are replaced by $\mathbf{x} - \bar{\mathbf{x}}$ and $\mathbf{X} - \bar{\mathbf{X}}$ respectively.

16.2.3 If a weighted regression is used, the expression for the leverage statistic becomes:

$$h = \mathbf{x^t(XRX^t)^+x} \qquad (66)$$

16.2.4 In MLR, if **m** is the vector (dimension $k$ by 1) of the selected values obtained from a spectral vector **x**, and **M** is the matrix of selected values for the calibration samples, then the leverage statistic is defined as:

$$h = \mathbf{m^t(MM^t)^{-1}m} \qquad (67)$$

16.2.5 Similarly, if a weighted regression is used, the expression for the leverage statistic becomes:

$$h = \mathbf{m^t(MRM^t)^{-1}m} \qquad (68)$$

16.2.6 In PCR and PLS, the leverage statistic for a sample with spectrum **x** is obtained by substituting the decompositions for PCR, or for PLS, into Eq 65. The statistic is expressed as:

$$h = \mathbf{s^t s} \qquad (69)$$

NOTE 16—If the scores from the PCR or PLS model are not normalized, then the form of Eq 69 becomes $\mathbf{D^2 = s^t(S^tS)^{-1}s}$

16.2.7 If a weighted PCR or PLS regression is used, the expression for the leverage statistic becomes

$$h = \mathbf{s^t(S^tRS)^{-1}s} \qquad (70)$$

16.3 *Outlier Detection During Calibration:*

16.3.1 Two types of outliers can be identified during the calibration procedures. The first type of outlier is a sample that represents an extreme composition relative to the remainder of the calibration set. These samples have very high leverage on the regression results; that is, they are largely responsible for the determination of at least one of the regression coefficient values. Generally, there is insufficient data in the calibration set to statistically determine the accuracy of reference values associated with these high leverage samples. Their inclusion in the calibration may lead to erroneous estimations of similar samples if the reference value for the high leverage sample is in error. The second type of outlier is one for which the estimated value differs from the reference value by a statistically significant amount. Such outliers indicate either an error in the reference measurement, an error in the spectral measurement, a clerical error in sample attribution or reference value transcription, or a failure of the model.

16.3.2 High-leverage samples are identified based on the leverage statistic, $h$ (also known as Mahalanobis Distance). For all types of linear calibrations described above, the average leverage statistic for all of the calibration sample spectra has a value of $k/n$ where $k$ is the number of variables in the regression (the number of wavelengths in MLR, the number of principal components, or the number of PLS latent variables), and $n$ is the number of calibration samples. On average, each sample contributes $k/n$ of the spectral variables. For samples that have $h > 3k/n$, the sample spectrum is contributing a significant fraction to the definition of one of the spectral variables and to the regression coefficient associated with this variable. Samples with $h > 3k/n$ should be carefully reviewed and considered for elimination from the calibration set in the development of the model. Often such high leverage samples will unduly dominate the determination of the regression coefficients to the detriment of model performance. However, in some instances a high leverage sample may contribute important information that can improve model performance, especially in cases where the signal to noise of the spectra is proportional to the leverage of the samples. Leverage is a valuable indicator of samples that may be detrimental to the calibration, however, the best measure of whether or not a sample should be eliminated is how well the resulting calibration can be validated.

NOTE 17—If the leverage statistic is scaled as described in (**25**), an $f$ test can be employed for outlier detection.

16.3.3 If calibration spectra with $h > 3k/n$ are eliminated from the calibration set, and the model is rebuilt, it is not uncommon for additional spectra with $h > 3k/n$ to be identified for the new model. This occurrence is most likely if removal of samples reduces $k$, but can also be caused merely by scaling changes to the multivariate space induced by changes in $n$.

When repetitive application of the 3k/n rule continues to identify outliers, the outlier test is said to "snowball." If "snowballing" occurs, it may indicate some problem with the structure of the spectral data set. The variable space of the model should be examined for unusual distributions or clusterings.

16.3.3.1 If the following sequence occurs during the development of a model, the 3k/n outlier test can be relaxed: (1) a first model is built on an initial calibration set, (2) calibration spectra with h >3k/n are eliminated from the calibration set, (3) a second model using the same number, k, variables is built on the subset of calibration spectra, and (4) calibration spectra with h >3k/n are identified for the second model. The second model should be used providing that no calibration samples have h greater than 0.5.

16.3.3.2 If (1) a first model is built on an initial calibration set, (2) calibration spectra with h >3k/n are eliminated from the calibration set, and (3) a second model using fewer variables is built on the subset of calibration spectra, the 3k/n outlier test should not automatically be relaxed. Instead, the first model should be rebuilt using the lower number of variables and the sequence in 16.3.3.1 should be applied to the new model.

16.3.4 A second type of outlier is one for which the estimated value ŷ differs by a statistically significant amount from the value from the reference method, y. Such outliers can be detected based on studentized residuals. If $e_i$ is the difference between the estimated value $\hat{y}_i$ and the reference value $y_i$ for the $i^{th}$ sample in the calibration set, and $h_i$ is the leverage statistic for that sample, the studentized residuals for the $i^{th}$ sample are given by:

$$t_i = \frac{e_i}{\text{SEC} \sqrt{1 - h}} \quad (71)$$

16.3.4.1 The studentized residuals should be normally distributed with common variance. The studentized residuals value can be compared to a t distribution value for $n - k$ (or $n - k - 1$ if mean centered) degrees of freedom, to determine the probability that the error in the estimate fits the expected distribution. If not, the sample should be considered an outlier. A more detailed discussion of studentized residuals can be found in Refs **26–27**.

16.3.5 If a sample is identified as an outlier based on studentized residuals or other similar tests, then the reference value may be in error. When possible, the reference test should be repeated to determine a correct value for the sample (multiple tests are recommended). If the reference value is not in error, then the large studentized residuals may indicate an error in the spectral measurement, a clerical error in sample attribution, or a basic failure in the model. For estimation of component concentrations, there may be sufficient spectral interferences to preclude accurate estimation of the component for this class of samples. For property estimation, some component that has a significant effect on the property may not be detected. Removing outliers of this type without evidence of error in the reference value should be avoided whenever possible, since these samples may provide the only indication that the model is not applicable to a certain class of materials.

16.4 *Interpolation and Extrapolation of the Model During Analysis:*

16.4.1 The spectra of the calibration samples define a set of variables that are used in the calibration of the multivariate model. If, when unknown samples are analyzed, the variables calculated from the spectrum of the unknown sample lie within the range of the variables for the calibration, the estimated value for the unknown sample is obtained by interpolation of the model. If the variables for the unknown sample are outside the range of the variables in the calibration model, the estimate represents an extrapolation of the model.

16.4.2 Two types of extrapolation are possible. First, the sample may contain the same components as the calibration samples, but at concentration ranges that are outside the ranges in the calibration set. Second, the sample may contain components that were not present in the calibration samples.

16.4.3 The leverage statistic, h, provides a useful indication of the first type of extrapolation. For the calibration set, one sample will have a maximum leverage statistic, $h_{max}$. This is the most extreme sample in the calibration set, in that, it is the farthest from the center of the space defined by the spectral variables. If the leverage statistic for an unknown sample is greater than $h_{max}$, then the estimate for the sample clearly represents an extrapolation of the model. Providing that outliers have been eliminated during the calibration, the distribution of h should be representative of the calibration model, and $h_{max}$ can be used as an indication of extrapolation.

NOTE 18—Comparison of the spectral variables for an unknown against the range of each spectral variable in the calibration model could be done, and extrapolation of any single variable could be taken as extrapolation of the model. The use of the leverage statistic as an indicator of extrapolation may not detect certain spectra which are slight extrapolations on one or more spectral variables; however, significant extrapolation of any one variable will result in a high leverage statistic, and thus detection of extrapolation. Use of individual variables in tests for extrapolation is not recommended since it can unduly restrict the range of samples to which the model is applicable.

16.4.4 The second type of extrapolation of the model, namely, the presence of a new component, can be detected by comparing an estimate of the unknown spectrum derived from the model to the measured spectrum of the unknown.

16.4.4.1 For PCR, an estimate of the spectrum of the unknown can be calculated as:

$$\mathbf{x^t} = \hat{\mathbf{s}}^t \sum \mathbf{L^t} \quad (72)$$

where the $\hat{\mathbf{s}}$ is the vector of scores. Similarly for PLS:

$$\mathbf{x^t} = \hat{\mathbf{s}}^t \mathbf{L^t} \quad (73)$$

where the $\hat{\mathbf{s}}$ is the vector of scores. The difference between the estimated spectrum and the actual spectrum can be calculated as:

$$\mathbf{r} = \mathbf{x} - \mathbf{x} \quad (74)$$

16.4.4.2 The root mean square spectral residuals (RMSSR) for the spectrum can then be calculated as:

$$\text{RMSSR} = \sqrt{\frac{\mathbf{r^t r}}{f}} \quad (75)$$

NOTE 19—Some commercial software packages may calculate other statistics related to RMSSR, or may call RMSSR by some other name. The model developer should verify what statistics are used in the software to indicate how well the model fits a spectrum being analyzed. The RMSSR

is intended as an example of how such a calculation can be done. Other similar statistics can be used.

16.4.5 The RMSSR values can be calculated for each of the calibration samples. One of the calibration samples will exhibit a maximum RMSSR, RMSSR$_{max}$. Assuming that outliers have been removed prior to the development of the calibration model, RMSSR$_{max}$ can be used to calculate a cutoff above which RMSSR values for unknown spectra are to be taken as evidence of extrapolation of the model.

16.4.6 In general, the RMSSR$_{max}$ cannot be used directly to set the cutoff for indicating extrapolation. For PCR and PLS models, some of the spectral noise characteristics of the calibration spectra are always incorporated into the spectral variables. The RMSSR values calculated for spectra used in the calibration will thus generally be lower than corresponding values calculated for spectra of the same samples which are not used in the model development. For estimating a suitable cutoff RMSSR value to serve as an indication of extrapolation, the following procedure is recommended.

16.4.6.1 Replicate spectral measurements (at least seven) of several (at least three) of the calibration samples should be made. The replicate measurements should include all steps in the measurement procedure (for example, background spectrum collection, loading of the sample, and measurement of the spectrum).

16.4.6.2 One spectrum from the set is to be used in the development of the calibration model. The RMSSR values for the spectra used in the calibration are calculated. The RMSSR$_{cal}(i)$ is the value for the spectrum of Sample $i$.

16.4.6.3 The remaining replicate spectra are analyzed using the calibration model, and RMSSR values are calculated and averaged for each sample. The RMSSR$_{anal}(i)$ is the average RMSSR for the replicate spectrum of Sample $i$.

16.4.6.4 The ratios of the RMSSR values from the analyses to those from the calibration are calculated and averaged, and RMSSR$_{max}$ is multiplied by the average ratio to obtain the cutoff:

$$RMSSR_{limit} = \frac{\left[ \sum_{i=1}^{m} \frac{RMSSR_{anal}(i)}{RMSSR_{cal}(i)} \right] RMSSR_{max}}{m} \qquad (76)$$

where $m$ is the number of replicate spectral measurements.

16.4.6.5 If the RMSSR value for an unknown sample being analyzed exceeds RMSSR$_{limit}$, then the analysis of the sample represents an extrapolation of the model.

16.4.7 Statistics comparable to RMSSR cannot be calculated for multiple linear regression. The MLR is thus incapable of detecting the second type of extrapolation, namely, the presence of a new component that was not in the calibration samples. Care should be exercised when applying MLR in systems where the calibration set used in the development of the MLR model may not represent the total range of sample compositions that will be encountered during analyses. In such cases, MLR should be supplemented with other techniques to determine if the sample being analyzed falls within the scope of the calibration. For example, outlier statistics from PCR models developed on the same calibration set could be used for this purpose.

NOTE 20—For PLS models, residuals calculations such as RMSSR are not always a useful indicator of outliers. If, during calibration, a significant percentage of the spectral(X-block) variance due to signal is not used in the model, then the model residuals used to calculate RMSSR$_{cal}$ may contain significant contributions due to calibration sample component absorptions. In such cases, RMSSR$_{limit}$ values calculated on the basis of such RMSSR$_{cal}$ values may be too large to detect model extrapolation due to new chemical components in samples being analyzed.

The procedure described in 15.3.3 can be used to estimate the percentage of the total X-block variance that is due to signal. If the variance included in the model is significantly less than the signal variance, then the modeler may wish to supplement the PLS model with a PCR model built on the same data. RMSSR statistics from the PCR model are then used for outlier detection. The number of variables used in the PCR model should be sufficient to account for the signal variance.

16.4.8 *Nearest Neighbor Distance*—If the calibration sample spectra form multiple clusters within the variable space, the spectrum of the unknown being analyzed can have a $D^2$ less than $D^2_{max}$ yet fall into a relatively unpopulated portion of the calibration space. In this case, the sample being analyzed contains the same components as the calibration samples (since the sample is not a RMSSR outlier), but at combinations that are not represented in the calibration set. The spectrum of the unknown does not belong to any of the calibration sample spectra clusters, and the results produced by application of the model may be invalid or less precise than the results for samples which do fall within these clusters. Under these circumstances, it is desirable to employ a Nearest Neighbor Distance test to detect unknown samples that fall within voids in the calibration space.

16.4.8.1 Nearest Neighbor Distance, NND, measures the distance between the spectrum being analyzed, **x**, and individual spectra in the calibration set, **x$_i$**.

$$NND = \min\left[ (\mathbf{x} - \mathbf{x_i})^t \, (\mathbf{XX^t})^{-1} \, (\mathbf{x} - \mathbf{x_i}) \right] \qquad (77)$$

16.4.8.2 For MLR, NND is calculated as

$$NND = \min\left[ (\mathbf{m} - \mathbf{m_i})^t \, (\mathbf{MM^t})^{-1} (\mathbf{m} - \mathbf{m_i}) \right] \qquad (78)$$

16.4.8.3 For PCR and PLS (with orthogonal scores), NND is calculated as

$$NND = \min\left[ (\mathbf{s} - \mathbf{s_i})^t \, (\mathbf{s} - \mathbf{s_i}) \right] \qquad (79)$$

16.4.8.4 NND values are calculated for all the calibration sample spectra. A maximum NND value is determined. This value represents the largest distance between calibration sample spectra.

16.4.8.5 During analysis, the NND value is calculated for the unknown sample spectrum relative to the calibration spectra. If the calculated value is greater than the maximum NND from 16.4.8.4, then the minimum distance between the process sample spectrum and the calibration spectra is greater than the largest distance between calibration sample spectra, the unknown sample spectrum falls within a sparsely populated region of the calibration space. Such samples are referred to as Nearest Neighbor Outliers.

## 17. Selection of Calibration Samples

17.1 For the development of a multivariate model, an ideal calibration sample set will:

17.1.1 Contain samples which provide examples of all chemical components which are expected to be present in the

samples which are to be analyzed using the model, thereby ensuring that analyses involve interpolation of the model;

17.1.2 Contain samples for which the range of variation in the concentrations of the chemical components exceeds the range of variation expected for samples which are to be analyzed using the model, thereby ensuring that analyses involve interpolation of the model;

17.1.3 Contain samples for which the concentrations of chemical components are uniformly distributed over their total range of variation.

Note 21—If it is desired to enhance the calibration model for unknown samples within particular area(s) of the calibration space this can be realized by including more calibration samples in the area(s) of the calibration space where enhancement is desired. Alternatively, it might be more appropriate to partition the calibration space in sub-spaces and develop separate calibrations for each sub space.

17.1.4 Contain a sufficient number of samples to statistically define the relationships between the spectral variables and the component concentrations or properties to be modeled.

17.2 For simple mixtures, calibration samples can generally be prepared to meet the criteria above. For complex mixtures, obtaining an ideal calibration set is difficult, if not impossible. The statistical tests that are used to detect outliers guard against non-ideal calibration sets. The RMSSR values detect when samples being analyzed contain components that are not represented in the calibration set (violation of 17.1.1). Leverage statistics detect when samples being analyzed are outside the concentration ranges represented in the calibration set (violation of 17.1.2). Outlier detection during model development identifies components for which the range of concentrations in the calibration set is not uniform (violation of 17.1.3).

17.3 The number of samples that are required to calibrate an infrared multivariate model (see 17.1.4) depends on the complexity of the samples being analyzed. If the samples to be analyzed contain only a few components that vary in concentration, then there will be a small number of spectral variables, and a relatively small calibration set is adequate to define the relationship between the variables and the concentrations or properties. If a larger number of components vary in the samples to be analyzed, then a larger number of calibration samples is required for the model development. Determining whether or not a set of calibration samples is adequate can only be done after a model is developed and an estimate of the number of spectral variables required for the model is made.

17.4 If a multivariate model is developed using three or fewer variables, then the calibration set should contain a minimum of 24 samples after elimination of outliers.

17.5 If a multivariate model is developed using $k$ ($>3$) variables, then the calibration set should contain a minimum of $6k$ spectra after elimination of outliers. If the model is mean centered, a minimum of $6(k + 1)$ spectra should remain.

Note 22—$6k$ is chosen to ensure at least 20 df in the model for statistical testing, and to ensure that there is an adequate number of samples to define the relationship between the spectral variables and the concentration or property values.

17.6 For some spectroscopic analyses, it is possible to calibrate using gravimetrically or volumetrically prepared mixtures which contain significantly fewer components than the samples which will ultimately be analyzed. For these surrogate methods, the outlier statistics described herein are not strictly appropriate since all actual samples are by definition outliers relative to the simplified calibrations. Thus, surrogate methods cannot strictly fulfill the requirements of this practice. Surrogate methods should, however, generally follow the requirements described herein for the number and range of calibration samples. In cases of highly linear systems where experimental design is used to construct the calibration space, surrogate calibrations may be successfully developed with fewer than the $6k$ samples normally required herein. Suitable experimental designs must ensure that all components are varied independently over an adequate range. Surrogate calibrations can be successfully deployed provided they are properly validated and are adequately supervised after deployment. However, surrogate methods may require significantly more validation as well as a heightened degree of supervision relative to non-surrogate methods.

## 18. Validation of a Multivariate Model

18.1 Validation of an infrared multivariate model is accomplished by applying the model for the analysis of a set of $v$ validation samples, and statistically comparing the estimates for these samples to known reference values. Validation requires thorough testing of the model to ensure that it performs up to the expectations derived from the calibration set statistics.

18.2 *Validation Sample Set:*

18.2.1 For the validation of a multivariate model, an ideal validation sample set will:

18.2.1.1 Contain samples that provide examples of all chemical components which are expected to be present in the samples which are to be analyzed using the model;

18.2.1.2 Contain samples for which the range of variation in the concentrations of the chemical components is comparable to the range of variation expected for samples that are to be analyzed using the model:

18.2.1.3 Contain samples for which the concentrations of chemical components are uniformly distributed over their total range of variation with particular attention to ensure that samples which may only occur rarely among all samples to be analyzed using the model are included in the validation set; and

18.2.1.4 Contain a sufficient number of samples to statistically test the relationships between the spectral variables and the component concentrations or properties that were modeled.

18.2.2 For simple mixtures, validation samples can generally be prepared to meet the criteria in 18.2.1.1 – 18.2.1.4. For complex mixtures, obtaining an ideal validation set is difficult if not impossible.

18.2.3 The number of samples needed to validate an infrared multivariate model depends on the complexity of the model. Only samples whose analyses are found to be interpolations of the model should be used in the validation procedure. If five or fewer spectral variables are used in the model, then a minimum of 20 interpolation samples is recommended. If $k >$ 5 spectral variables are used in the model, then a minimum of

$4k$ interpolation samples should be used in the validation. In addition, the validation samples should:

18.2.3.1 Span the range of concentrations or property values for which the model was developed; that is, the span and the standard deviation of the range of concentrations or property values for the validation samples should be at least 95 % of the span and the standard deviation of the range of concentrations or property values in the model, and the concentration or property values for the validation samples should be distributed as uniformly as possible across the range; and

18.2.3.2 Span the range of spectral variables for which the model was developed; that is, if the range of a spectral variable in the calibration model is from $a$ to $b$, and the standard deviation of the spectral variable is $c$, then the spectral variables for the validation samples should cover at least 95 % of the range from $a$ to $b$, and should be distributed as uniformly as possible across the range such that the standard deviation in the spectral variables estimated for the validation samples will be at least 95 % of $c$.

18.2.4 Determination of whether a validation set is adequate will generally require that the set be analyzed so that the spectral variables for the set can be determined. Samples whose analyses are extrapolations of the model should not be included in the validation set. If the validation set does not meet the criteria in 18.2.3.1 and 18.2.3.2, additional validation samples should be taken.

18.2.5 When feasible, it is useful to prepare a separate validation set comprising outlier samples. This outlier validation set can be used to test the effectiveness of any outlier detection which may be part of the calibration and to give an indication of the type and magnitude of problems which could be expected if an outlier (for which the calibration is, by definition invalid) should be accidently estimated by the calibration.

18.3 *Validation Spectra Measurement and Analysis*—Spectra of validation samples should be collected using exactly the same procedures as were used to collect spectra of the calibration samples. Reference values for the validation samples should be obtained using the same reference method as was used for the calibration samples. Spectra should be analyzed using the multivariate model to produce estimates of the component concentrations or properties, and the statistics described in Sections 18 and 19 should be calculated.

18.4 *Validation Error:*

18.4.1 If $v$ (a vector of dimensions $v$ by one) are the estimates obtained by analysis of the spectra of the $v$ validation samples, and $v$ are the corresponding values measured by the reference method, then the validation error, $\mathbf{e}$ is given by:

$$\mathbf{e} = \mathbf{v} - \mathbf{v} \qquad (80)$$

18.4.2 If multiple reference values are available for some of the validation samples, then the average of the individual reference measurements can be used in $\mathbf{v}$, and the variance removed by calculating the averages should be calculated using Eq 56.

18.5 *Variance of the Validation Error*—The variance of the error of the validation measurements is calculated as:

$$\text{VAR}_v = \mathbf{e}^t \mathbf{R} \mathbf{e} + \sigma_{avg}^2 = \sum_{i=1}^{v} \sum_{j=1}^{r_i} (\text{v}_{ij} - \text{v}_i)^2 \qquad (81)$$

where $\sigma_{avg}^2$ is zero and $R$ is an identity matrix if individual reference measurements are used in $v$.

18.6 *Standard Error of Validation:*

18.6.1 The standard error of validation (SEV) is given by:

$$SEV = \sqrt{\frac{\text{VAR}_v}{d_v}} = \sqrt{\frac{\sum_{i=1}^{v} \sum_{j=1}^{r_i} (\text{v}_{ij} - \text{v}_i)^2}{\sum_{i=1}^{v} r_i}} \qquad (82)$$

$d_v$ is the total number of reference values available for all $v$ validation samples. SEV is the standard deviation in the differences between reference and IR estimated values for samples in the validation set. The standard error of validation is sometimes referred to as a standard error of prediction or incorrectly as the standard error of estimate (SEE). A bias corrected version of this statistic has also been defined as the standard error of performance. To avoid confusion between two terms that are both abbreviated SEP, the use of SEV is preferred in these practices.

18.6.2 Studentized residuals testing can be applied to the estimates of the validation set to detect possible errors in the reference values.

18.7 *Validation Bias*—The average bias for the estimation of the validation set, $\bar{\text{e}}_v$, is calculated as:

$$\bar{\text{e}}_v = \frac{\sum_{j=1}^{v} r_i e_i}{d_v} = \frac{\sum_{i=1}^{v} \sum_{j=1}^{r_i} (\text{v}_{ij} - \text{v}_i)}{\sum_{i=1}^{v} r_i} \qquad (83)$$

where $r_i$ is 1 if individual reference values were used, or is the number of reference values that were averaged for the $i^{th}$ validation sample if averages are used. $d_v$ is the total number of reference values used in the calculation.

18.8 *Standard Deviation of Validation Errors*—The standard deviation of the validation errors, SDV, is calculated as

$$\text{SDV} = \sqrt{\frac{\sum_{i=1}^{v} r_i(\text{e}_i - \bar{\text{e}}_v)^2 + \sigma_{avg}^2}{d_v - 1}} = \sqrt{\frac{\sum_{i=1}^{v} \sum_{j=1}^{r_i} [(\text{v}_{ij} - \text{v}_i) - \bar{\text{e}}_v]^2}{\left(\sum_{i=1}^{v} r_i\right) - 1}} \qquad (84)$$

where $r_i$ is 1 and $\sigma_{avg}^2$ is 0 if individual reference measurements are used in calculating $\hat{\mathbf{y}}$.

18.9 *Significance of Validation Bias:*

18.9.1 A $t$ test is used to determine if the validation estimates show a statistically significant bias. A $t$ value is calculated as:

$$t = \frac{|\bar{e}_v| \sqrt{d_v}}{SDV} \qquad (85)$$

The $t$ value is compared to critical $t$ values from Table A1.3 for $d_v$ degrees of freedom.

18.9.2 If the $t$ value is less than the critical $t$ value, then analyses based on the multivariate model are expected to give essentially the same average result as measurements conducted by the reference method, provided that the analysis represents an interpolation of the model.

18.9.3 If the *t* value calculated is greater than the tabulated *t* value, there is a 95 % probability that the estimate from the multivariate model will not give the same average results as the reference method. Validity of the multivariate model is then suspect. Further investigation of the model is required to resolve the probable bias that is indicated.

18.10 *Validation of Agreement Between Model and Reference Method:*

18.10.1 The confidence limits on the estimates for the validation samples should be calculated, and a determination should be made as to whether the individual reference values for the validation samples lie within the range from $\hat{y} - t \times SEC \times \sqrt{1 + D^2}$ to $\hat{y} + t \times SEC \times \sqrt{1 + D^2}$. If more than 5 % of the reference values fall outside this range, then the confidence limit estimates based on SEC are questionable, and further testing is required to demonstrate the agreement between the model and the reference method.

18.10.2 An alternative method can be used to demonstrate agreement between the model and the reference method. This alternative method is preferred when the precision of the reference method is not constant across the range of reference values used in the calibration, but can be applied even when the precision is constant. If $R(y_i)$ is the reproducibility of the reference method at level $y_i$, then the percentage of reference values for which:

$$\hat{y}_i - R(\hat{y}_i) < y_{ij} < \hat{y}_i + R(\hat{y}_i) \tag{86}$$

is calculated. If 95 % or more of the reference values fall within this interval, then estimates produced with the multivariate IR model agree with those produced by the reference method as well as a second laboratory repeating the reference measurement would agree.

18.11 For multivariate analyses employing surrogate calibrations, a procedure similar to that described here for validation is often performed for the purpose of verifying that the instrument is properly calibrated. This instrument qualification procedure typically involves the analysis of gravimetrically or volumetrically prepared mixtures that contain significantly fewer components than the samples which will ultimately be analyzed. There is no a priori relationship between the standard error that is calculated from this procedure and the error expected during application of the model to actual samples. To avoid confusion, it is recommended that the procedure be referred to as a spectrometer/spectrophotometer qualification, not validation. Additionally, it is recommended that the standard error calculated from this procedure be referred to as a Standard Error of Qualification ($SEQ_{surrogate}$), not as a Standard Error of Validation.

## 19. Precision of Infrared Estimated Values

19.1 The precision of values estimated from an infrared multivariate model is calculated from repeated spectral measurements. The number of samples for which repeat measurements is made should be at least equal to the number of variables used in the model, and never less than three. The samples used for repeat spectral measurements should span at least 95 % of the range of concentration or property values used in the model. When possible, samples should be selected to ensure that some variation on each spectral variable is exhibited among the samples. At least six spectra should be collected for each sample. The spectra should be analyzed and values estimated. The average estimate for each sample should be calculated, and the standard deviation among the estimates should be obtained. If $y_{ij}$ is the estimate for the $j^{th}$ spectrum of $r_i$ total spectra for the $i^{th}$ sample, then the average estimate for this sample is:

$$\bar{\hat{y}}_i = \frac{\sum_{j=1}^{r_i} \hat{y}_{ij}}{r_i} \tag{87}$$

19.1.1 The standard deviation of the replicate estimates is calculated as:

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{r_i} (\hat{y}_{ij} - \bar{y}_i)^2}{r_i - 1}} \tag{88}$$

19.2 A $\chi^2$ value is calculated using the standard deviation values calculated in Eq 85:

$$\chi^2 = \frac{2.3026}{c} \left( r \log\sigma^2 - \sum_{i=1}^{r_i} r_i \log\sigma_i^2 \right) \tag{89}$$

where:

$$r = \sum_{i=1}^{t} r_i \tag{90}$$

$$\sigma = \sqrt{\frac{1}{r} \sum_{i=1}^{t} r_i \sigma_i^2} \tag{91}$$

$$c = 1 + \frac{1}{3(z-1)} \left( \sum_{i=1}^{z} \frac{1}{r_i} - \frac{1}{r} \right) \tag{92}$$

and $z$ is the number of samples for which replicate measurements were made.

19.3 The $\chi^2$ value calculated in Eq 89 is compared with a critical value from a chi-squared table (see Table A1.4) for $t - 1$ degrees of freedom. If the calculated $\chi^2$ value is less than the critical value, then all of the variances for the replicated measurements belong to the same population, and the average variance calculated in Eq 91 can be used as a measure of the repeatability of the infrared measurement. The infrared analysis is expected to have a repeatability on the order of $t \times \sqrt{2}\bar{\sigma}$.

19.4 If the calculated $\chi^2$ value is greater than the critical chi-squared value, then the repeatability of the infrared estimate may vary with sample composition. In this case, the infrared analysis is expected to have a repeatability that is no worse than $t \times \sqrt{2} \times \sigma_{max}$, where $\sigma_{max}$ is the maximum $\sigma_i$ value for the replicate measurements.

## 20. Major Sources of Calibration and Analysis Error

20.1 *General Sources of Error in Spectral Measurements*— Table 1 list some possible sources of error that can occur during the spectral measurement and potential solutions for these problems.

20.2 *Sampling Related Errors*—Table 2 lists errors arising from sampling problems and possible solutions to these problems (28).

20.3 *Sources of Calibration Error*—Table 3 lists sources of error in the development of the calibration model and possible ways to minimize these errors.

**TABLE 1 General Sources of Error in Spectral Measurements**

| Source of Spectral Error | Possible Solution |
|---|---|
| Poor instrument performance | Conduct instrument performance tests regularly to monitor changes in instrument performance |
| | Analyze QC (Quality Check) sample to determine if instrument performance changes affect analysis |
| Absorbance exceeds linear response range | Determine linear response range for instrument |
| | Choose pathlengths to keep bands of interest in range |
| Optical polarization effects | Use depolarizing elements |
| Variable sample presentation | Improve sample presentation methods |
| | Investigate commercially available sample presentation equipment |
| Optical component contamination | Inspect windows, etc., for contamination and clean as necessary |

**TABLE 2 Sampling Related Errors**

| Sampling Error | Possible Solution |
|---|---|
| Nonhomogeneity of sample | Improve mixing guidelines or grinding procedures, or both |
| | For solids, average replicate repacks |
| | For solids, rotate sample cups |
| | Measure multiple aliquots from large sample volume |
| Physical variation in solid samples | Improved sample mixing during sample preparation |
| | Diffuse light before it strikes the sample using a light diffusing plate |
| | Pulverize sample to particle size of less than 40 μm (NIR) or 2 μm (MIR) |
| | Average multiple repacks of each sample |
| | Rotate sample or average five sample measurements |
| Chemical variation in sample with time | Freeze-dry sample for storage and measurement |
| | Immediate data collection and analysis following sample preparation |
| | Identification of kinetics of chemical change and avoidance of rapidly changing spectral regions |
| Bubbles in liquid samples | Check pressure requirements for single-phase sample |
| | Check flow properties of cell for sample introduction |

**TABLE 3 Sources of Calibration Error**

| Source of Calibration Error | Possible Solution |
|---|---|
| Spectroscopy insensitive to component/property being modeled | Try alternative spectral region |
| | Redefine requirement in terms of measurable components/properties |
| Inadequate sampling of population in calibration set | Review criteria for calibration set selection |
| | Use sample selection techniques for selecting calibration set (29) |
| Outlier samples within calibration set | Employ outlier detection algorithms |
| | Eliminate spectral outliers or find additional examples |
| | Eliminate reference data outliers or remeasure |
| Reference data errors | Analyze blind replicates to test precision |
| | Correct procedural errors, improve analytical procedures |
| | Check and recalibrate reagents, equipment, etc. (30) |
| Non-Beer's Law relationship (Nonlinearity due to component interactions) | Develop multiple calibrations over smaller concentration ranges |
| (Nonlinearity due to instrument response) | Check dynamic range of instrument, Try shorter pathlengths |
| Sensitivity to baseline shifts, etc. | Preprocessing of data to minimize effects of baseline |
| Transcription errors | Two people cross-check or one person triple-check all handscribed data |

**TABLE 4 Analysis Errors**

| Sources of Analysis Error | Possible Solution |
|---|---|
| Poor calibration model | Validate calibration model on representative validation set |
| Poor instrument performance | Check performance of instrument/model with QC samples |
| | Diagnose instrument problems with instrument performance tests |
| Poor calibration transfer | Validate calibration transfer and instrument standardization procedures |
| | Select calibrations with lowest noise, wave length shift sensitivity, and offset sensitivity |
| Sample outside model range | Employ outlier statistics to test that sample is interpolation of model |

20.4 *Analysis Errors*—Table 4 lists factors that can contribute to errors in the estimated values for unknown samples and possible ways to minimize such errors.

## 21. Wavelength (Frequency) Sensitivity of a Multivariate Model

21.1 Wavelength stability of spectrometers is often a critical factor in the performance of a multivariate calibration. The estimation of the sensitivity of a multivariate model to changes in the wavelength scale provides a useful parameter against which instrument performance can be judged. The wavelength sensitivity of a model can be roughly estimated by the following procedure:

21.1.1 Identify the samples in the calibration set that represent the extreme values of each of the spectral variables;

21.1.2 If the spectra are collected with a digital resolution of $\Delta$, then shift each spectrum by $+\Delta$ and by $-\Delta$.

21.1.3 Analyze the shifted spectra, using the calibration model, and calculate the change in the estimates between the $+\Delta$ and $-\Delta$ spectra, and

21.1.4 Identify the spectrum showing the largest change upon shifting. If the estimates are $\hat{\mathbf{y}}_{+\Delta}$ and $\hat{\mathbf{y}}_{-\Delta}$ respectively, then the wavelength (frequency) sensitivity of the model can be estimated as:

$$0.1 \times \Delta \times \text{SEC}/(\hat{\mathbf{y}}_{+\Delta} - \hat{\mathbf{y}}_{-\Delta}) \qquad (93)$$

21.2 The value calculated in Eq 93 is the wavelength shift that, in the worst case (the most sensitive spectrum) will produce a change in the estimate that is on the order of 5 % of the standard error of calibration.

NOTE 23—The wavelength sensitivity of a model calculated in Section 12 will depend on a variety of factors, including the optical and digital resolution of the instrument relative to the bandwidths of the sample being measured. Calculation of a wavelength sensitivity is done to provide a useful diagnostic for analyses conducted on the same type of analyzer. The wavelength stability of the analyzer can be compared to the value in Eq 87 as a means of monitoring the performance of the analyzer. Because the value in Eq 87 is dependent on specific instrumental parameters, it should generally not be used to compare the suitability of analyzers for a particular application.

## 22. Calibration Transfer and Instrument Standardization

22.1 Calibration transfer refers to a process by which a calibration model is developed using data from one or more spectrometers, is possibly modified, and is applied for the

analysis of spectra collected on additional spectrometers. The calibration transfer may require that spectral data for a common sample or samples be collected on both instruments, and that some transfer function be developed and applied to the spectra or the model or the estimate produced by the model. A complete description of calibration transfer methodologies is beyond the scope of these practices.

22.2 Instrument standardization is a process where the spectra collected on a second instrument are mathematically adjusted in an attempt to match the spectra that would have been collected on the instrument on which the calibration was developed. Instrument standardization can also involve actual adjustment of the instrument hardware to achieve such agreement. Instrument standardization is one means of achieving calibration transfer.

22.3 Calibration transfer or instrument standardization may be required when maintenance is done to an instrument if such maintenance produces a change in the spectral response large enough to change the values estimated by the calibration model. The calibration can be thought of as being transferred from one instrument (before maintenance) to a second instrument (after maintenance).

22.4 When a calibration transfer or instrument standardization procedure is developed, it is necessary to demonstrate that the performance of the model is not degraded during the transfer. To demonstrate that a calibration transfer or instrument standardization procedure preserves the performance of a model, it is necessary to validate the transferred model as described in Section 18. Each calibration transfer or instrument standardization procedure must be tested at least once by performing a full validation of the transferred model. Once the success of a particular calibration transfer or instrument standardization procedure has been demonstrated for a particular type of instrument, then quality control samples can be used to evaluate additional transfers and standardizations.

## 23. Calibration Quality Control

23.1 When an IR, multivariate, analysis is used to estimate component concentrations or properties, or both, it is desirable to periodically test the analysis (instrument and model) to ensure that the performance of the analysis is unchanged. To perform such tests, it is sometimes necessary to choose one or more quality control samples that will be used for this purpose. A complete discussion of methods used to validate the performance of an IR analyzer is beyond the scope of these practices. The user is referred to Practice D6122 which discusses validation of IR analyzers for hydrocarbon analysis, and to Refs 30 and 31 which discuss methods that have gained acceptance within the agricultural community.

23.2 Control samples (materials for which reference values have been measured using the reference method) can be employed to monitor the performance of the analysis, provided that the analyses of the control samples involve interpolation of the model. The IR estimated values for the control samples are compared to the reference values using established ASTM procedures or alternative statistical tests (30, 32). These tests will generally require that the IR estimated values and the reference values agree to within the confidence intervals defined in 15.3. Since the confidence limits are based on SEC, and since SEC is often dominated by the error in the reference measurement, these procedures may not provide the most sensitive indication of changes in the performance of the analysis. Alternatively, quality control (QC) samples can be employed.

23.3 Quality control (QC) samples are used to monitor changes in the performance of an analysis (instrument and model), after the analysis has been validated. Quality control materials should be identified at the time the model is developed based on the following criteria:

23.3.1 QC materials must be chemically and physically compatible with materials being analyzed, so as to not introduce contaminants into the samples being analyzed, and not to cause safety problems.

23.3.2 QC materials must be chemically stable when stored and sampled. If mixtures are used, the composition of the mixture must be known and methods for reproducing the mixture must be established.

23.3.3 The spectra of the QC material must be compatible with the model. Absorption bands for the QC material should not exceed the linear response range of the instrument in regions used in the calibration model. The spectra of the QC material should be as similar as possible to spectra of the calibration samples. However, analysis of the QC sample can be an extrapolation of the model.

23.4 Spectral data on the QC material is collected during the same time period that spectra of the calibration and validation samples are collected. The QC material should be treated in exactly the same fashion as other samples so that variations in the spectra are representative of the variations which will occur during the collection of spectra for unknowns. Separate samples should be used for each measurement. A minimum of 20 spectra should be collected.

NOTE 24—If the QC spectra are collected over too short a time interval, the variation seen in the spectra will be smaller than that typically encountered in application of the model to unknowns, and QC limits set based on these spectra will be excessively tight.

23.4.1 The spectra for the QC material are analyzed using the calibration model, and the average value, $\bar{y}_{qc}$ is calculated:

$$\bar{y}_{qc} = \frac{\sum_{i=1}^{q} \hat{y}_i}{q} \tag{94}$$

where $q$ is the number of spectra collected for the QC material. The standard deviation in the estimated values, $\sigma_{qc}$, is calculated as

$$\sigma_{qc} = \sqrt{\frac{\sum_{i=1}^{q} (\hat{y}_i - \bar{y}_{qc})^2}{q - 1}} \tag{95}$$

23.4.1.1 Dixon's test can be applied to the individual estimated values to identify outliers in the calculations in Eq 94 and Eq 95.

23.5 The QC material is analyzed periodically when the analysis (instrument and model) is in use for analyzing unknowns. The QC material is treated exactly the same as an

unknown sample being estimated. The estimated value for the QC material is compared to $y_{qc}$. The estimated value is expected to be within the range from $y_{qc} - t \times \sigma_{qc}$ to $y_{qc} + t \times \sigma_{qc}$ 95 % of the time, where $t$ is the studentized $t$ value for q − 1 df and the 95 % confidence level.

23.5.1 If the analysis of the QC material is an interpolation of the model, then $\sigma_{qc}$ should be consistent with the repeatability of the IR analysis as defined in Section 19. If the analysis of the QC material is an extrapolation of the model, then $\sigma_{qc}$ may be somewhat higher than the $\sigma_i$ calculated in Section 19. However, since the control limits are still based on the repeatability of the spectral measurement and do not depend on the reference method, they are expected generally to be tighter than those derived from control samples.

23.6 The use of bias and slope adjustments to improve calibration or prediction statistics for IR multivariate models is appropriate provided that the calibration is fully revalidated whenever bias and slope is adjusted. Bias and slope adjustments are another form of calibration transfer and use of bias and slope adjustments should be handled in the same fashion as any other calibration transfer. Prediction errors requiring continued bias and slope corrections indicate drift in reference method or changes in the character of the samples, sample handling, sample presentation, instrument response function, or wavelength stability. If a calibration model fails during the QC monitoring step, the performance of the instrument should be evaluated using the appropriate ASTM instrument performance test, and any instrument problem that is identified should be corrected. If control samples are used, checks should be performed on the reference method to ensure that reference values are correct. If instrument maintenance is performed, calibration transfer or instrument standardization procedures, or both, should be followed to reestablish the calibration.

## 24. Model Updating

24.1 It may sometimes be desirable to add additional calibration samples to an existing model to increase the range of applicability of the model. The new calibration samples may contain the same components as the original calibration samples but at more extreme concentrations, or new components not present in the original calibration samples. The new calibration samples may fill voids in the original calibration space.

24.1.1 When a model is updated, the matrix **X** containing the original calibration spectra is augmented with the spectra of the additional calibration samples, and the vector **y** containing the property or composition values for the calibration samples is augmented with the values for the additional calibration samples.

24.1.2 Outlier procedures described in 16.3 must be applied to updated models in the same way they are applied to new models. Thus, if additional samples are being added to increase the span of the calibration, it may be necessary to add several samples of each new type to avoid the added samples being rejected as outliers.

24.2 When a calibration model is updated, it must be revalidated. The requirements for validation samples for an updated model are the same as for the original model (see

Section 18). The spectra used to validate the original model can be used to validate the updated model, but they must be supplemented to cover an adequate range as described in 18.2. The percentage of new samples added to the validation set for the updated model must be at least as large as the percentage of new samples added to the calibration set.

## 25. Multivariate Calibration Questionnaire

25.1 The following questionnaire is designed to assist the user in determining if a multivariate calibration conforms to the requirements set forth in these practices.

25.1.1 If all of the following questions in 25.1.3 – 25.1.7 are answered in the affirmative, then the calibration can be said to have been developed and validated according to E1655.

25.1.2 If any of the following questions in 25.1.3 – 25.1.7 are answered in the negative, then the calibration can not be said to have been developed and validated according to E1655. If the calibration method is MLR, PCR or PLS-1, the calibration may be said to have been developed using mathematical techniques described in E1655. ASTM methods that reference E1655 should not claim calibration or validation via E1655 unless all of the following questions would have been answered in the affirmative for the procedures followed during the collection of round robin data on which the method is based.

25.1.3 The following questions apply to the mathematical methodology used in the calibration:

25.1.3.1 Was the mathematical technique used in the calibration MLR, PCR or PLS-1? (Sections 12 and 13)

25.1.3.2 Did the calibration methodology include the capability of detecting high leverage outliers using a statistic such as the leverage statistic, $h$? (16.2)

25.1.3.3 Did the analysis methodology include the capability to detect outliers via a statistic such as those based on spectral residuals? (16.4.4 – 16.4.7)

25.1.4 The following questions apply to the calibration model where $n$ is the number of samples in the calibration set, and $k$ is the number of variables (MLR wavelengths, Principal Components, or PLS latent variables) in the model.

25.1.4.1 Was $n > 6k$ if the model is not mean centered, or $n > 6(k + 1)$ if the model is mean centered? (17.5)

25.1.4.2 Was the number of samples in the calibration set at least 24? (17.4)

25.1.5 The following questions apply to the validation of the model:

25.1.5.1 Was a separate set of validation samples used to test the calibration? (18.2)

25.1.5.2 Were validation spectra which were outliers based on either leverage (Mahalanobis Distance) or spectral residuals excluded from the validation set? (18.2.3)

25.1.5.3 Was the number of validation samples greater than $4k$ if the model was not mean centered, or greater than $4(k + 1)$ if the model was mean centered? (18.2.3)

25.1.5.4 Was the number of validation samples at least 20? (18.2.3)

25.1.5.5 Did the validation samples span 95 % of the range of the calibration samples? (18.2.3.1)

25.1.5.6 If SEC is the Standard Error of Calibration, do 95 % of the results for the validation samples fall within $\pm$ t·SEC·$\sqrt{1+h}$ of the reference values where t is the Studentized $t$ value for $n-k$ degrees of freedom ($n-k-1$ for mean centered models), and $h$ is the leverage statistic? (18.10.1)

25.1.5.7 Do the validation results show a statistically insignificant bias? (18.9.1)

25.1.6 Was the precision of the model determined using $t \geq k \geq 3$ test samples and $r \geq 6$ replicate measurements per sample? (Section 19)

25.1.7 If the calibration and analysis methodology includes preprocessing or postprocessing, are these calculations performed automatically? (Sections 11 and 14)

## 26. Keywords

26.1 infrared analysis; molecular spectroscopy; multivariate analysis; quantitative analysis

## ANNEXES

### (Mandatory Information)

### A1. STATISTICAL TREATMENT

#### A1.1 Dixon's Test Functions for Rejection of Outliers

A1.1.1 This test provides a simple and highly efficient method for determining whether all data obtained came from the same population (with unknown mean and standard deviation) and if one or more of the data points are suspect and should be rejected.

A1.1.2 In applying this test the number of determinations ($N$) are tabulated in increasing order of magnitude and designated as $X_1, X_2, X_3, \ldots X_n$.

A1.1.3 The values at the extremes of the tabulation $X_1$ and $X_n$ are tested in turn in accordance with the number of values in the tabulation.

A1.2 Select the proper expression shown as follows in accordance with the number ($N$) of the values in the tabulation and the upper or lower limit to be tested:

| Outliers Under Test For $N =$ | $X_1$ | $X_n$ |
|---|---|---|
| 3 to 7 | $r = \dfrac{(X_2 - X_1)}{(X_n - X_1)}$ | $r = \dfrac{(X_n - X_{(n-1)})}{(X_n - X_1)}$ |
| 8 to 10 | $r = \dfrac{(X_2 - X_1)}{(X_{(n-1)} - X_1)}$ | $r = \dfrac{(X_n - X_{(n-1)})}{(X_n - X_2)}$ |
| 11 to 13 | $r = \dfrac{(X_3 - X_1)}{(X_{(n-1)} - X_1)}$ | $r = \dfrac{(X_n - X_{(n-2)})}{(X_n - X_1)}$ |
| 14 to 30 | $r = \dfrac{(X_3 - X_1)}{(X_{(n-2)} - X_1)}$ | $r = \dfrac{(X_n - X_{(n-2)})}{(X_n - X_3)}$ |

A1.3 Substitute the appropriate values in the equation selected, calculate "$r$" and compare the value obtained to the $r$ value in Table A1.1 for the appropriate sample size ($N$).

A1.4 Reject the value if the calculated "$r$" is greater than the tabulated value.

**TABLE A1.1 Critical Values for Rejection of a Discordant Measurement (31)**

| Statistic | N | $\alpha = 0.05$ | $\alpha = 0.01$ |
|---|---|---|---|
| $r_{10}$ | 3 | 0.941 | 0.988 |
| | 4 | 0.765 | 0.889 |
| | 5 | 0.642 | 0.780 |
| | 6 | 0.560 | 0.698 |
| | 7 | 0.507 | 0.637 |
| $r_{11}$ | 8 | 0.554 | 0.683 |
| | 9 | 0.512 | 0.635 |
| | 10 | 0.477 | 0.597 |
| $r_{21}$ | 11 | 0.576 | 0.679 |
| | 12 | 0.546 | 0.642 |
| | 13 | 0.521 | 0.615 |
| $r_{22}$ | 14 | 0.546 | 0.641 |
| | 15 | 0.525 | 0.616 |
| | 16 | 0.507 | 0.595 |
| | 17 | 0.490 | 0.577 |
| | 18 | 0.475 | 0.561 |
| | 19 | 0.462 | 0.547 |
| | 20 | 0.450 | 0.535 |
| | 21 | 0.440 | 0.524 |
| | 22 | 0.430 | 0.514 |
| | 23 | 0.421 | 0.505 |
| | 24 | 0.413 | 0.497 |
| | 25 | 0.406 | 0.489 |

A1.5 Historical standard deviation as used in Fig. A1.1 means the standard deviation of a test method. It is established by averaging the standard deviations of many samples tested by many laboratories. The samples should cover the range of usefulness of the test method and should include materials of diverse composition if the latter has any effect on the reproducibility of results.

A1.6 *Sample Standard Deviation* is merely the standard deviation computed from the data obtained by a group of laboratories testing the same sample using the same test method. Obviously it may be much lower or much higher than the historical standard deviation of the test method. Therefore the *sample* standard deviation may be less reliable (because of

**TABLE A1.2 *F*-Distribution: Degrees of Freedom for Numerator**

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161 | 200 | 216 | 225 | 230 | 234 | 237 | 239 | 241 | 242 | 244 | 246 | 248 |
| 2 | 18.5 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 |
| 3 | 10.1 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.87 | 8.85 | 8.81 | 8.79 | 8.74 | 8.70 | 8.66 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.91 | 5.86 | 5.80 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.06 | 4.95 | 4.88 | 4.81 | 4.77 | 4.74 | 4.68 | 4.62 | 4.56 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.00 | 3.94 | 3.87 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.57 | 3.51 | 3.44 |
| 8 | 5.32 | 4.46 | 4.07 | 3.54 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.28 | 3.22 | 3.15 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.07 | 3.01 | 2.94 |
| 10 | 4.96 | 4.10 | 3.70 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.91 | 2.85 | 2.77 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.79 | 2.72 | 2.55 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.69 | 2.62 | 2.54 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.60 | 2.53 | 2.46 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.53 | 2.46 | 2.39 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.48 | 2.40 | 2.33 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.42 | 2.35 | 2.28 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.38 | 2.31 | 2.23 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.34 | 2.27 | 2.19 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.31 | 2.23 | 2.16 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.28 | 2.20 | 2.12 |
| ∞ | 3.84 | 3.00 | 2.50 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 | 1.83 | 1.75 | 1.67 | 1.57 |

**TABLE A1.3 Table of *t* at 5 % Probability Level**

| Degrees of Freedom | *t* |
|---|---|
| 1 | 12.706 |
| 2 | 4.303 |
| 3 | 3.182 |
| 4 | 2.776 |
| 5 | 2.571 |
| 6 | 2.447 |
| 7 | 2.365 |
| 8 | 2.306 |
| 9 | 2.262 |
| 10 | 2.228 |
| 11 | 2.201 |
| 12 | 2.179 |
| 13 | 2.160 |
| 14 | 2.145 |
| 15 | 2.131 |
| 16 | 2.120 |
| 17 | 2.110 |
| 18 | 2.101 |
| 19 | 2.093 |
| 20 | 2.086 |

**TABLE A1.4 Critical $\chi^2$ Values**

NOTE 1— $\chi^2$ values for $(t - 1)$ degrees of freedom and 95 % confidence level.

| $(t-1)$ | $\chi^2$ | $(t-1)$ | $\chi^2$ | $(t-1)$ | $\chi^2$ | $(t-1)$ | $\chi^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 3.84 | 6 | 12.59 | 11 | 19.68 | 16 | 26.30 |
| 2 | 5.99 | 7 | 14.07 | 12 | 21.03 | 17 | 27.59 |
| 3 | 7.81 | 8 | 15.51 | 13 | 22.36 | 18 | 28.87 |
| 4 | 9.49 | 9 | 16.92 | 14 | 23.68 | 19 | 30.14 |
| 5 | 11.07 | 10 | 18.31 | 15 | 25.00 | 20 | 31.41 |

| No. of Results | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| Factor | 2.20 | 1.62 | 1.42 | 1.31 | 1.25 |
| No. of Results | 8 | 10 | 15 | 25 | 35 |
| Factor | 1.21 | 1.15 | 1.09 | 1.05 | 1.04 |

A1.7 *To Find the Number of Determinations Needed in an Average to Give Specific Confidence Limits*—Lay a straightedge across the nomograph so that its edge passes through the point on the right scale corresponding to the standard deviation for the test and through the desired point on the confidence limit scale. Read the number of determinations required from the left scale.

A1.8 *To Find the Confidence Limits of an Average*—Using the number of determinations in the average, lay a straightedge from this point on the left scale through the point on the right scale corresponding to the standard deviation. Read the confidence limits from the intermediate scale.

these random fluctuations) than the *historical* standard deviation in determining the confidence limits of an average of results of several determinations.

A1.6.1 If the historical standard deviation is unknown, the sample standard deviation may be substituted for it in using the nomograph and then multiplying the value found on the 95 % CL scale by the factor given as follows for the number of results in the average to obtain reliable 95 % confidence limits.
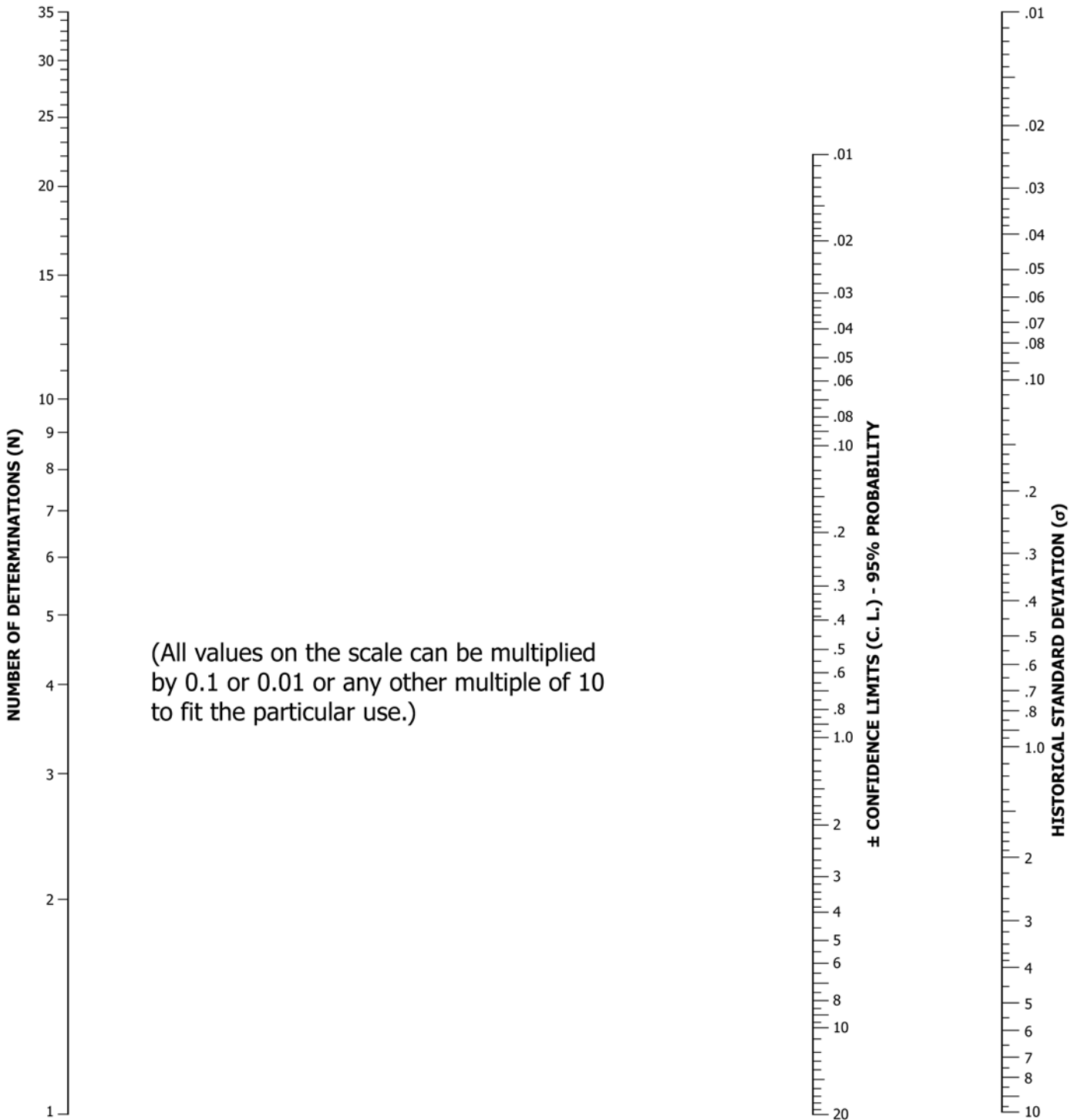
FIG. A1.1 Nomograph for Number of Determinations to Obtain Desired Confidence Limits

(All values on the scale can be multiplied by 0.1 or 0.01 or any other multiple of 10 to fit the particular use.)

**A2. STATISTICAL TESTS COMMON TO NIRS METHODS (18, 19) (SUPPLEMENTAL INFORMATION)**

**A2.1 Common Symbols**

A2.1.1 Throughout these practices, lowercase letters are used to represent scalar quantities. Lower case **bold** letters are used to represent vectors, and upper case **BOLD** letters are used to represent matrices. Italicized letters are used to represent dimensions of vectors and matrices. Italicized subscripts are sample, wavelength indices. For example:

$y_i$ = Scalar reference value for the $i^{th}$ sample.
$\hat{y}_i$ = The estimated y-value for $i^{th}$ sample based on a regression model.
$\bar{y}$ = The mean y value for all samples.
**y** = Vector of reference values for $n$ samples.
$\mathbf{x}_i$ = Spectral vector of length $f$ for the $i^{th}$ sample.
**X** = Matrix of spectra, the $n$ rows of **X** contain the spectra of length $f$ for $n$ samples.
$n$ = Number of samples used in a calibration model.
$f$ = Number of frequencies or wavelengths used in a calibration model.
$k$ = Number of variables used in a calibration model.
$r$ = Number of replicate measurements on a sample.
$\sum$ = Capital sigma represents summation of all values within parentheses.
$\mathbf{R}^2$ = Coefficient of multiple determination (R-squared).
$\mathbf{R}$ = The simple correlation coefficient for a linear regression for any set of data points; this is equal to the square root of the R-squared value.
$b_0$ = The bias or y-intercept value for any calibration function fit to x, y data. For bias-corrected standard error calculations the bias is equal to the difference between the average reference analytical values and the IR predicted values.

**A2.2 Statistical Terms**

A2.2.1 Sum of squares for regression:

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad (A2.1)$$

A2.2.2 Sum of squares for residual:

$$SS_{res} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \qquad (A2.2)$$

A2.2.3 Mean square for regression:

$$MS_{reg} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{k - 1} \qquad (A2.3)$$

A2.2.4 Mean square for residual:

$$MS_{reg} = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n - k - 1} \qquad (A2.4)$$

A2.2.5 Total sum of squares:

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (A2.5)$$

**A2.3 Test Statistics**

A2.3.1 The statistics discussed as follows have most commonly been applied to MLR models. The statistics assume that the data has been mean centered in developing the model. Similar statistics can be derived for PCR and PLS models, and for models that are not mean centered.

A2.3.2 *Coefficient of Multiple Determination*—The coefficient of multiple determination is also termed the R-squared statistic, or total explained variation. This statistic allows determination of the amount of variation in the data that is adequately modeled by the calibration equation as a total fraction of 1.0. Thus $R^2 = 1.00$ indicates the calibration equation models 100 % of the variation within the data. An $R^2 = 0.50$ indicates that 50 % of the variation in the differences between the actual values for the data points and the predicted or estimated values for these points are explained by the calibration equation (mathematical model), and 50 % is not explained. Squared values approaching 1.0 are attempted when developing calibrations. R-squared can be estimated using a simple method as outlined as follows.

A2.3.2.1 The $R^2$ is determined using the equation:

$$\mathbf{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)/(n - k - 1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2/(n - 1)} = \frac{SS_{reg}}{SS_{tot}} \qquad (A2.6)$$

A2.3.2.2 If $\sigma_R$ is the standard deviation of the errors in the reference method measurement, and $\sigma_Y$ is the standard deviation in the reference values used in the calibration (a measure of the range spanned by the reference data), then $R^2$ values that exceed $1 - \sigma_R^2/\sigma_Y^2$ are probable indications of overfitting of the data.

A2.3.3 *F-Test Statistic for the Regression:*

A2.3.3.1 This statistic is also termed F for regression, or t-squared. F increases as the equation begins to model, or fit, more of the variation within the data. With R-squared held constant, the F value increases as the number of samples increases. As the number wavelengths used within the regression equation decreases, F tends to increase. Deleting an unimportant wavelength from an equation will cause the F for regression to increase.

A2.3.3.2 The F-statistic can also be useful in recognizing suspected outliers within a calibration sample set; if the F-value decreases when a sample is deleted, the sample was not an outlier. This situation is the result of the sample not affecting the overall fit of the calibration line to the data while at the same time decreasing the number of sample ($n$). Conversely, if deleting a single sample increases the overall F for regression, the sample is considered a suspected outlier. F is defined as the mean square for regression divided by the mean square for residual (see statistical terms in A1.2).

A2.3.3.3 The F for the regression is determined by the equation:

$$F = \frac{R^2(n-k-1)}{(1-R^2)k} = \frac{MS_{reg}}{MS_{res}} \quad (A2.7)$$

A2.3.4 *Student's t-Value (For a Regression):*

A2.3.4.1 This statistic is equivalent to the F statistic in the determination of the correlation between **X** and **y** data. It can be used to determine whether there is a true correlation between an IR estimated value and the primary chemical analysis for that sample. It is used to test the hypothesis that the correlation really exists and has not happened only by chance. A large t value (generally greater than ten) indicates a real (statistically significant) correlation between **X** and **y**.

A2.3.4.2 The t for regression is calculated as:

$$t = \frac{R\sqrt{n-k-1}}{\sqrt{1-R^2}} \quad (A2.8)$$

A2.3.5 *Partial F or t-Squared Test for a Regression Coefficient:*

A2.3.5.1 This test indicates whether the addition of a particular wavelength (independent variable) and its corresponding regression coefficient (multiplier) adds any significant improvement to an equation's ability to model the data (including the remaining unexplained variation). Small F or t values indicate no real improvement is given by adding the wavelength into the equation.

A2.3.5.2 If several wavelengths (variables) have low t or F values (less than 10 or 100, respectively), it may be necessary to delete each of the suspect wavelengths, singly or in combination, to determine which wavelengths are the most critical for predicting constituent values. In the case where an important wavelength is masked by intercorrelation with another wavelength, a sharp increase in the partial F will occur when an unimportant wavelength is deleted and where there is no longer high intercorrelation between the variables still within the regression equation.

A2.3.5.3 The t-statistic is sometimes referred to as the ratio of the actual regression coefficient for a particular wavelength to the standard deviation of that coefficient. The partial F value described is equal to this t value squared; note that the t value calculated this way retains the sign of the coefficient, whereas all F values are positive.

A2.3.5.4 The partial F for a regression coefficient is calculated as:

$$\frac{SS_{res}(\text{all variables except one}) - SS_{res}(\text{all variables})}{MS_{res}(\text{all variables})} \quad (A2.9)$$

A2.3.6 *The Bias Corrected Standard Error:*

A2.3.6.1 Bias corrected standard error measurements allow the characterization of the variance attributable to random unexplained error within. The bias value, $b_0$, is calculated as the mean difference between reference and IR estimated values:

$$b_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (A2.10)$$

A2.3.6.2 The bias corrected standard error is calculated as:

$$SE_c = \sqrt{\frac{\sum_{i=1}^{n}((y_i - \hat{y}_i) - b_0)^2}{n-1}} \quad (A2.11)$$

Similar bias corrected values can be calculated for SECV.

A2.3.7 *Standard Deviation of Repeatability (SDR):*

A2.3.7.1 SDR is also referred to as the standard deviation of difference (SDD) or standard error of difference for replicate measurements (SD replicates). The SDR is calculated to allow accurate estimation of the variation in an analytical method due to both sampling, sample presentation, and analysis errors. The SDR can be used as a measure of precision for the reference analytical method.

A2.3.7.2 The SDR is calculated using:

$$SDR = \sqrt{\frac{\sum_{j=1}^{r}(y_j - \bar{y}_j)^2}{r-1}} \quad (A2.12)$$

A2.3.8 *Offset Sensitivity:*

A2.3.8.1 Also termed systematic variation or index of systematic variation (ISV), offset sensitivity is equal to the sum of all regression coefficients. The larger the value, the greater is the sensitivity to particle size differences between samples or to the isotropic (mirror-like) scattering properties of samples. The offset sensitivity is used to compare two or more equations for their "blindness" to offset variation between samples. Equations with large offset sensitivities indicate that particle size variations within a data set may cause wide variations in the analytical result.

A2.3.8.2 The ISV is calculated as:

$$ISV = \sum_{i=1}^{k} b_i \quad (A2.13)$$

A2.3.9 *Random Variation Sensitivity:*

A2.3.9.1 This statistic is also termed the index of random variation (IRV). Random variation sensitivity is calculated as the sum of the squares of the values of all regression coefficients. The larger the value, the greater the sensitivity to factors such as: poor wavelength precision, temperature variations within samples and instrument, and electronic noise. The higher the value, the less likely the equation can be transferred successfully to other instruments.

A2.3.9.2 The IRV is calculated using the expression:

$$IRV = \sum_{i=1}^{k} \sqrt{b_i^2} \quad (A2.14)$$

A2.3.10 *Standard Error of the Laboratory (SEL) for Reference Chemical Methods:*

A2.3.10.1 The SEL can be determined by using one or more samples properly aliquoted and analyzed in replicate by one or more laboratories. The average analytical value for the replicates on a single sample is determined as:

$$\bar{y}_i = \sum_{j=1}^{r} y_{ij} \quad (A2.15)$$

A2.3.10.2 SEL is given by:

$$SEL = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n(r_i - 1)}} \quad (A2.16)$$

where the *i* index represents different samples and the *j* index different measurements on the same sample.

A2.3.10.3 This can apply whether the replicates were performed in a single laboratory or whether a collaborative study was undertaken at multiple laboratories. Additional techniques for planning collaborative tests can be found in Ref **20**. Some care must be taken in applying Eq. 2.3.16. If all of the analytical results are from a single analyst in a single laboratory, then the repeatability of the analysis is defined as $\sqrt{2}$ t($n(r - 1)$, 95 %) SEL, where t($n(r - 1)$, 95 %) is the Student's $t$ value for the 95 % confidence level and $n(r - 1)$

degrees of freedom. If the analytical results are from multiple analysts and laboratories, the same calculation yields the reproducibility of the analysis. For many analytical tests, SEL may vary with the magnitude of y. SEL values calculated for samples having different $\bar{y}_i$ can be compared by an F-test to determine if the SEL values show a statistically significant variation as a function of $\bar{y}_i$.

## REFERENCES

**(1)** Association of Official Analytical Chemists, *AOAC Official Methods of Analysis*, Method 989.03, 1990, pp. 74–76.

**(2)** *Journal of the Association of Official Analytical Chemists*, Vol 71, 1988, p. 1162.

**(3)** Landa, I., *Review of Scientific Instruments*, Vol 50, 1979, pp. 34–40.

**(4)** Landa, I., and Norris, K. H., *Applied Spectroscopy*, Vol 23, 1979, pp. 105–107.

**(5)** Kortüm, G., *Reflectance Spectroscopy*, Springer-Verlag, New York, NY, 1969, p. 111.

**(6)** Honigs, D. E., Freelin, J. M., Hieftje, G. M., and Hirschfeld, T. B., *Applied Spectroscopy*, Vol 37, No. 6, 1983, pp. 491–497.

**(7)** Hrushka, W., "Data Analysis: Wavelength Selection Techniques," in *Near Infrared Technology in the Agricultural and Food Industries*, P. Williams and K. Norris, Eds., American Association of Cereal Chemists, St. Paul, MN, 1987.

**(8)** Mark, H., *Applied Spectroscopy*, Vol 42, No. 8, 1988, pp. 1427–1440.

**(9)** Brown, P. J., *Journal of Chemometrics*, Vol 6, 1992, pp. 151–161.

**(10)** Fredricks, P. M., Osborn, P. R., and Swinkels, P. R., *Analytical Chemistry*, Vol 57, 1985, pp. 1947–1950.

**(11)** Kennedy, W. J., and Gentle, J. E., *Statistical Computing*, Marcel Dekker, New York, NY, 1980.

**(12)** Allen, D. M., *Technical Report Number 23*, University of Kentucky Department of Statistics, August 1981.

**(13)** Lindberg, W., Persson, J., and Wold, S., *Analytical Chemistry*, Vol 55, 1983, p. 643.

**(14)** Martens, H. A., and Naes, T., *Multivariate Calibration*, John Wiley and Sons, New York, NY, 1989.

**(15)** Geladi, P., and Kowalski, B. R., *Journal of Chemometrics*, Vol 1, 1986, pp. 1 and 18.

**(16)** Haaland, D. M., and Thomas, E. V., *Analytical Chemistry*, Vol 60, 1988, pp. 1193–1202.

**(17)** Wold, S., Ruhe, A., Wold, H., and Dunn, W. J., *SIAM Journal of Science and Statistical Computations*, Vol 5, 1984, p. 735.

**(18)** Manne, R., *Chemometrics and Intelligent Laboratory Systems*, Vol 2, 1987, p. 187.

**(19)** Helland, I. S., *Communications in Statistics (Simulation and Computation)*, Vol 17, 1988, p. 581.

**(20)** Helland, I. S., *Scandinavian Journal of Statistics*, Vol 17 , 1990, p. 97.

**(21)** Draper, N. R., and Smith, A., *Applied Regression Analysis*, John Wiley and Sons, New York, NY, 1981.

**(22)** Workman, J., "NIR Spectroscopy Calibration Basics," in *Near Infrared Analysis*, Burns, D., and Ciurczak, E., eds., Marcel-Dekker, Inc., New York, NY, 1992, pp. 247–280.

**(23)** Mark, H., and Workman, J., *Statistics in Spectroscopy*, Academic Press, Boston, MA, 1991.

**(24)** Hoaglin, D.C., Welsch, R.E. *Amer. Statist.* 1978 , 32, 17.

**(25)** Whitfield, R.G., Gerger, M.E., and Sharp, R.L., *Applied Spectroscopy*, Vol 41, 1987, pp. 1204–1213.

**(26)** Geladi, P., and Kowalski, B. R., Analytica Chimica Acta, 185, 1986, pp. 1–17.

**(27)** Miller, R., *Simultaneous Inference*, 2nd ed., Springer, New York, NY, 1981.

**(28)** Mark, H., and Workman, J., *Analytical Chemistry*, Vol 58, 1986, p. 1454.

**(29)** Honigs, D. E., Hieftje, G. M., Mark, H. L., and Hirschfeld, T. B., *Analytical Chemistry*, Vol 57, 1985, p. 2299.

**(30)** Youden, W. J., *Statistical Manual of the Association of Official Analytical Chemists*, AOAC, Arlington, VA, 1979.

**(31)** Martens, H., and Naes, T., In Williams, P., and Norris, K., Eds., *Near Infrared Technology in the Agricultural and Food Industries*, American Association of Cereal Chemists, St. Paul, MN, 1987, pp. 57–87.

**(32)** Hald, A., *Statistical Theory with Engineering Applications*, John Wiley and Sons, New York, NY, 1952.

**(33)** Dixon, W. J., *Biometrics*, Vol 9, 1953, pp. 74–89.