Appendix K: Guidelines for Dietary Supplements and Botanicals

This appendix contains three complementary documents for the validation of dietary supplements and botanical methods:

Part I: AOAC Guidelines for Single-Laboratory Validation of Chemical Methods for Dietary Supplements and Botanicals

Part II: AOAC Guidelines for Validation of Botanical Identification Methods

Part III: Probability of Identification: A Statistical Model for the Validation of Qualitative Botanical Identification Methods

PART I
AOAC Guidelines for Single-Laboratory Validation of Chemical Methods for Dietary Supplements and Botanicals

Contents

Under a 5-year contract (2003–2008) with the National Institutes of Health-Office of Dietary Supplements, through the U.S. Food and Drug Administration, AOAC undertook an effort to validate methods for dietary supplement ingredients of interest. As part of the initiative, AOAC adapted and revised the traditional *Official Methods*℠ process to include single-laboratory validation (SLV). Methods were first validated within a single laboratory to test their suitability and ruggedness without the complications of a multilaboratory collaborative study. SLVs proved to be an excellent debugging tool for complex methods; problems found within one laboratory could be dealt with so that a stronger method went on to the collaborative study. The SLV process, thus, became a step in preparation for the collaborative study.

The SLV guidelines were approved by the AOAC Official Methods Board and Board of Directors in December 2002.

Because of the time and expense required for the determination of modern analytes such as pesticide residues, industrial contaminants, veterinary drugs, allergens, botanicals, dietary supplements, and alternative medicines in complex matrices, there is considerable interest in obtaining acceptable methods of analysis faster and cheaper. It has been suggested that accreditation of laboratories, internal quality control, and external proficiency exercises can improve laboratory performance to the point where interlaboratory validation is no longer an absolute necessity. To this end AOAC INTERNATIONAL has been exploring alternatives to the full interlaboratory study design that requires the examination of a minimum of five matrices by eight laboratories (*see* www.aoac.org under method validation programs). These have included "minicollaborative" studies that reduced the required number of matrices and laboratories, the "Peer-Verified Methods Program," which merely required verification of the analytical parameters by a second laboratory, "*Performance Tested Methods*^{SM}" for test kits, the developing e-CAM compiling program (www.AOAC.org/AOAC_e-CAM.pdf), and the International Union of Pure and Applied Chemistry (IUPAC) sanctioned single-laboratory validation (SLV) protocol [*Pure & Appl. Chem*. **74**(5), 835–855(2002)].

The IUPAC single-laboratory protocol necessarily deals in generalities and specifically points out, "The total cost to the analytical community of validating a specific method through a collaborative trial and then verifying its performance attributes in the laboratories wishing to use it, is frequently less than when many laboratories all independently undertake SLV of the same method." The protocol also indicates that the degree of validation depends upon the status of the method in the analytical structure. At one extreme is the initial application of a well-established method in a laboratory that merely requires verification of the capability of that laboratory to achieve the published performance characteristics. The opposite extreme is the initial presentation of a new method or the initial application of an established method to a new matrix or application. Methods that are developed in response to a continued need for compliance, surveillance, and enforcement of laws and contracts involving a number of laboratories are expected to proceed to a multilaboratory validated status.

This AOAC document is intended to present guidelines for the evaluation of the initial use of a new or old method in a laboratory. It assumes that a proposed or available method is fairly well developed, optimized, and stabilized, that it has been applied to some practical test samples with acceptable results, and that a description of the method and its initial performance results are available in some kind of document. The initiating or another laboratory must then decide if the demonstrated performance appears to be satisfactory for the same or for another purpose.

Although the output from method development is the input to method validation, method developers cannot expect much input from method validators. Although method validators may have had considerable experience in the analysis of practical analytical samples, they are not expected to have the basic knowledge to recommend improvement in methods, such as certain solvents as useful for extraction of certain classes of analytes or column-solvent combinations as useful for optimization of separations. Method developers are expected to bring methods to the point where they satisfy validation requirements.

By definition, SLV does not provide any information on what values would be expected on examination of identical test samples by other laboratories. Therefore such methods probably would be used by regulatory agencies only for monitoring purposes—to explore compliance with laws and regulations unless the statutes under which they operate assign correctness to their results. Ordinarily such methods would not be used to bring a legal action or to settle a commercial dispute until their properties had been further explored in an environment provided by an interlaboratory collaborative study or a proficiency study utilizing that method. As stated in the FDA Center for Drug Evaluation and Research (CDER) "Reviewer Guidance/Validation of Chromatographic Methods" (November 1994), "Methods should be reproducible when used by other analysts, on other equivalent equipment, on other days and locations, and throughout the life of the drug product."

1 Introduction

The primary purpose of validating a method of analysis is to show that the method is fit for its intended purpose. Some purposes are:

(*1*) Determine how much of a valuable, necessary, or characteristic ingredient is present in a product.

(*2*) Determine if a product meets specifications.

(*3*) Determine if a product meets regulatory requirements.

(*4*) Survey an environment to determine the presence and amount of a component, contaminant, or a nutrient.

(*5*) Identify a product and/or its components.

The purposes usually answer the questions, "What is this product?" in the sense of its common or usual name, chemical identity, or components, and "How much of something [an analyte] is in this product [matrix]?"

At least at the initial stages of a problem, only a single or at most a very few laboratories require validation of a method of analysis. These circumstances include situations similar to the following:

(*1*) Methods for research.

(*2*) Only a few test samples are anticipated.

(*3*) For quality control of a manufacturing process of a single item by a single producer,

(*4*) Checking the reliability of a method imported from another source.

(*5*) Rechecking the reliability of a previously used method after a period of disuse.

(*6*) Situations where there is a lack of interest by other laboratories in participating in an interlaboratory validation exercise.

(*7*) Multi-analyte, multi-matrix methods where a conventional interlaboratory validation exercise is impractical.

For the present purpose we assume:

(*1*) We know or can assume the chemical identity of the material we are dealing with.

(*2*) We have a specimen of the material that can be used as a reference to compare the signal produced by the analyte isolated from the product we are examining with the same signal produced by a known amount of the reference analyte (traceable to a stated reference).

If either or both of these requirement are not met, much useful information can still be obtained, but our information will be "floating" in the same sense as a ship at sea does not know where it is without landmarks to determine its position. If the identity of an analyte must

be determined, not merely verified, a whole new dimension is added to the problem. This involves bringing in a laboratory or an individual with skill in determining chemical structure, a highly specialized, expensive, and time-consuming exercise.

It is often found during the initial experience with application or validation of a method that deficiencies appear, unexpected interferences emerge, reagents and equipment are no longer available, instruments must be modified, and other unanticipated problems require returning the method to a development phase. Frequently a method that functions satisfactorily in one laboratory fails to operate in the same manner in another. Often there is no clear-cut differentiation between development and validation and the two procedures constitute an iterative process. For that reason some aspects of method development that provide an insight into method performance, such as ruggedness, are included in this document.

In some cases it is impossible to set specific requirements because of unknown factors or incomplete knowledge. In such cases it is best to accept whatever information is generated during development and validation and rely upon the "improvements" that are usually forthcoming to asymptotically approach performance parameters developed for other analytes in the same or in a similar class.

1.1 Definitions

1.1.1 Validation

Validation is the process of demonstrating or confirming the performance characteristics of a method of analysis.

This process of validation is separate from the question of acceptability or the magnitude of the limits of the characteristics examined, which are determined by the purpose of the application. Validation applies to a specific operator, laboratory, and equipment utilizing the method over a reasonable concentration range and period of time.

Typically the validation of a chemical method of analysis results in the specification of various aspects of reliability and applicability. Validation is a time-consuming process and should be performed only after the method has been optimized and stabilized because subsequent changes will require revalidation. The stability of the validation must also be verified by periodic examination of a stable reference material.

1.1.2 Method of Analysis

The method of analysis is the detailed set of directions, from the preparation of the test sample to the reporting of the results, that must be followed exactly for the results to be accepted for the stated purpose.

The term "method of analysis" is sometimes assigned to the technique, e.g., liquid chromatography or atomic absorption spectrometry, in which case the set of specific directions is referred to as the "protocol."

1.1.3 Performance Characteristics of a Method of Analysis

The performance characteristics of a method of analysis are the functional qualities and the statistical measures of the degree of reliability exhibited by the method under specified operating conditions.

The functional qualities are the selectivity (specificity), as the ability to distinguish the analyte from other substances; applicability, as the matrices and concentration range of acceptable operation; and degree of reliability, usually expressed in terms of bias as recovery, and variability as the standard deviation or equivalent terms (relative standard deviation and variance).

Measurements are never exact and the "performance characteristics of a method of analysis" usually reflect the degree to which replicate measurements made under the same or different conditions can be expected or required to approach the "true" or assigned values of the items or parameters being measured. For analytical chemistry, the item being measured is usually the concentration, with a statement of its uncertainty, and sometimes the identity of an analyte.

For abbreviations and symbols used in this guideline, *see Annex A*.

2 Single-Laboratory Validation Work

2.1 Preparation of the Laboratory Sample

Product and laboratory sampling are frequently overlooked aspects of analytical work because very often product sampling is not under the control of the laboratory but the sample is supplied by the customer. In this case, the customer assumes the responsibility of extrapolating from the analytical result to the original lot. If the laboratory is requested to sample the lot, then it must determine the purpose of the analysis and provide for random or directed sampling accordingly.

The laboratory is responsible for handling the sample in the laboratory to assure proper preparation with respect to composition and homogeneity and to assure a suitable analytical sample. The laboratory sample is the material received by the laboratory and it usually must be reduced in bulk and fineness to an analytical sample from which the test portions are removed for analysis.

Excellent instructions for this purpose will be found in the "Guidelines for Preparing Laboratory Samples" prepared by the American Association of Feed Control Officials, Laboratory Methods and Service Committee, Sample Preparation Working Group (2000) (AAFCO, Oxford, IN) that cover the preparation of particularly difficult mineral and biological material. The improper or incomplete preparation of the analytical sample is an often overlooked reason for the nonreproducibility of analytical results.

If a laboratory prepares test samples for the purpose of validating a method, it should take precautions that the analyst who will be doing the validation is not aware of the composition of the test samples. Analysts have a bias, conscious or unconscious, of permitting knowledge of the identity or composition of a test sample to influence the result [*J. AOAC Int.* **83**, 399–406(2000)].

2.2 Identification

Identification is the characterization of the substance being analyzed, including its chemical, mineral, or biological classification, as applicable. In many investigations the identity of the analyte is assumed and the correctness of the assumption is merely confirmed. With some products of natural origin, complete identification and characterization is not possible. In these cases identification often may be fixed by chemical, chromatographic, or spectrophotometric fingerprinting—producing a reproducible pattern of reactions or characteristic output signals (peaks) with respect to position and intensity.

For botanical products, provide:

- Common or usual name of the item
- Synonyms by which it is known
- Botanical classification (variety, species, genus, family)

- Active or characteristic ingredient(s) (name and Chemical Abstracts Registry number or Merck Index number) and its chemical class. If the activity is ascribable to a mixture, provide the spectral or chromatographic fingerprint and the identity of the identifiable signals.

2.3 Method of Analysis or Protocol

The protocol or method of analysis is the set of permanent instructions for the conduct of the method of analysis. The method of analysis that is finally used should be the same as the one that was studied and revised as a result of research, optimization, and ruggedness trials and edited to conform with principles and practices for the production of *Official Methods of Analysis of AOAC INTERNATIONAL* (OMA). At this point the text is regarded as fixed. Substantive changes (those other than typographical and editorial) can only be made by formal public announcement and approval.

This text should be in ISO-compatible format where the major heads follow in a logical progression [e.g., Title, Applicability (Scope), Equipment, Reagents, Text, Calculations, with the addition of any special sections required by the technique, e.g., chromatography, spectroscopy]. Conventions with respect to reagents and laboratory operations should follow those given in the section "Definition of Terms and Explanatory Notes," which explains that "water is distilled water," reagents are of a purity and strength defined by the American Chemical Society (note that these may differ from standards set in other parts of the world), alcohol is the 95% aqueous mixture, and similar frequently used working definitions.

AOAC-approved methods may be considered as "well-recognized test methods" as used by ISO 17025. This document requires that those method properties, which may be major sources of uncertainties of measurements, be identified and controlled. In AOAC methods the following operations or conditions, which may be major contributors to uncertainties, should be understood to be within the following limits, unless otherwise specified more strictly or more loosely:

- Weights: Within ±10% (but use actual weight for calculations)
- Volumes: Volumetric flasks, graduates, and transfer pipets (stated capacity with negligible uncertainty)
- Burets: Stated capacity except in titrations
- Graduated pipets: Use volumes >10% of capacity
- Temperatures: Set to within ±2°
- pH: Within ±0.05 unit
- Time: Within ±5%

If the operational settings are within these specifications, together with any others derived from the supporting studies, the standard deviation obtained from these supporting studies in the same units as the reported result with the proper number of significant figures, usually 2 or 3, may be used as the standard measurement uncertainty.

2.3.1 Optimization

Prior to determining the performance parameters, the method should be optimized so that it is fairly certain that the properties of the "final method" are being tested. Validation is not a substitute for method development or for method optimization. If, however, some of the validation requirements have already been performed during the development phase, there is no need to repeat them for the validation phase. A helpful introduction is the AOAC publication "Use of Statistics to Develop and Evaluate Analytical Methods" by Grant T. Wernimont. This volume has only three major chapters: the measurement process, intralaboratory studies, and interlaboratory studies. No simpler explanation in understandable chemical terms exists of the analysis of variance than that given in pages 28–31. It supplements, explaining in greater detail, the concepts exemplified in the popular "Statistical Manual of AOAC" by W.J. Youden. Other useful references are *Appendices D* and *E* of OMA.

2.3.2 Reference Standard

All chemical measurements require a reference point. Classical gravimetric methods depend on standard weights and measures, which are eventually traceable to internationally recognized (SI) units. But modern analytical chemistry depends on other physical properties in addition to mass and length, usually optical or electrical, and their magnitude is based upon an instrumental comparison to a corresponding physical signal produced from a known mass or concentration of the "pure" analyte. If the analyte is a mixture, the signals or components must be separated and the signal from each compound compared to the signal from a known mass or concentration of the pure material or expressed in terms of a single reference compound of constant composition.

All instrumental methods require a reference material, even those that measure an empirical analyte. An "empirical analyte" is an analyte or property whose value is not fixed as in stoichiometric chemical compounds but which is the result of the application of the procedure used to determine it; examples are moisture, ash, fat, carbohydrate (by difference), and fiber. It is a "method-dependent analyte." Usually the reference material or "standard," which are specific chemical compounds, can be purchased from a supplier of chemicals and occasionally from a national metrological institute. When used for reference purposes, a statement should accompany the material certifying the identity, the purity and its uncertainty, how this was measured (usually by spectroscopy or chromatography), and its stability and storage conditions. If no reference material is available, as with many isolates from botanical specimens, an available compound with similar properties may serve as a surrogate standard—a compound that is stable and which behaves like the analyte but which is well resolved from it. Sometimes an impure specimen of the analyte must serve temporarily as the reference material until a purer specimen becomes available. The measured values assigned to empirical analytes are determined by strict adherence to all the details of the method of analysis. Even so, their bias and variability are usually larger (poorer) than chemically specified analytes. In some cases, as in determining the composition of milk by instrumental methods, the reference values for fat, protein, and lactose are established by use of reference methods. In routine operation, the bias and uncertainty of the final values are the combination of the uncertainties and bias correction arising from the routine operation with that of the reference values used for the calibration.

Modern instrumentation is complicated and its operation requires training and experience not only to recognize acceptable performance but also to distinguish unacceptable performance, drift, and deterioration on the part of the components. Continuous instruction and testing of the instruments and operators with in-house and external standards and proficiency exercises are necessary.

The records and report must describe the reference material, the source, and the basis for the purity statement (certification by the supplier is often satisfactory). If the reference material is hygroscopic, it should be dried before use either in a 100°C oven, if stable, or over a drying agent in a desiccator if not. The conversion factor of the analyte to the reference material, if different, and its

uncertainty must be established, often through spectrophotometric or chromatographic properties such as absorptivity or peak height or area ratios.

For recovery experiments the reference standard should be the highest purity available. In the macro concentration range (defined as about 0.1–100%) the standard ordinarily approaches 100%; in the micro or trace (defined as μg/g to 0.1%) and ultramicro or ultratrace range (μg/g and below) the standard should be at least 95% pure. The purity of rare or expensive standards is often established, referenced, and transferred through an absorptivity measurement in a specific solvent. The impurities present should not interfere significantly with the assay.

2.3.3 Ruggedness Trial

Although the major factors contributing to variability of a method may be explored by the classical, one variable at a time procedure, examining the effect of less important factors can be accomplished by a simpler Youden Ruggedness Trial [Youden, W.J., & Steiner, E.H. (1975) *Statistical Manual of the Association of Official Analytical Chemists*, pp 50–55]. This design permits exploring the effect of 7 factors in a single experiment requiring only eight determinations. It also permits an approximation of the expected standard deviation from the variability of those factors that are "in control." An example of exploring the extraction step of the determination of the active ingredient in a botanical is detailed in *Annex B*.

2.3.4 Specific Variables

If a variable is found to have an influence on the results, further method development is required to overcome the deficiency. For example, extraction of botanicals is likely to be incomplete and there are no reference materials available to serve as a standard for complete extraction. Therefore various techniques must be applied to determine when extraction is complete; reextraction with fresh solvent is the most common. Considerable experimentation also may be necessary to find the optimum conditions, column, and solvents for chromatographic isolation of the active ingredient(s).

(**a**) *Analyte addition*.—Addition of a solution of the active ingredient to the test sample and conducting the analysis is generally uninformative because the added analyte is already in an easily extractable form. The same is true for varying the volume of the extracting solvent. These procedures do not test the extractability of the analyte embedded in the cell structure. For this purpose, other variables must be tried, such as changing the solvent polarity or the extraction temperature.

(**b**) *Reextraction of the extracted residue*.—Reextraction after an original extraction will test for complete extraction by the original procedure. It will not test for complete extraction from intractable (unextractable) plant material. For this purpose a reagent that will destroy fibrous cellular material without damaging the active ingredient is required. If the analytes will not be destroyed or interfered with by cell wall disrupting or crude fiber reagents (1.25% H_2SO_4 and 1.25% NaOH) and are water soluble, use these solutions as extractives. But since the active ingredients are likely to contain compounds hydrolysable by these reagents, mechanical grinding to a very fine mesh will be the more likely choice.

The efficiency of extraction is checked by application of the extract to TLC, GLC, or HPLC chromatography. Higher total extractables is not necessarily an indicator of better extraction. The quantification of the active ingredient(s) is the indicator of extraction. Many natural compounds are sensitive to light and the decrease of a component suggests that the effect of this variable should be investigated.

(**c**) *Comparison with different solvents*.—Solvents with different polarities and boiling points will extract different amounts of extractives, but the amount of active ingredient(s) must be pursued by chromatographic separation or by specific reactions.

(**d**) *Comparison with results from a different procedure*.—A number of analyte groups, e.g., pesticide residues, have several different standard methods available based on different principles to provide targets for comparison.

(**e**) *System suitability checks*.—Chromatographic systems of columns, solvents (particularly gradients), and detectors are extremely sensitive to changes in conditions. Chromatographic properties of columns change as columns age and changes in polarity of solvents or temperature must be made to compensate. Therefore the specified properties of chromatographic systems in standard methods such as column temperatures and solvent compositions are permitted to be altered in order to optimize and stabilize the chromatographic output—peak height or area, peak resolutions, and peak shape. Similarly optical filters, electrical components of circuits, and mechanical components of instruments deteriorate with age and adjustments must be made to compensate. Specifications for instruments, and their calibration and operation must be sufficiently broad to accommodate these variations.

3 Performance Characteristics

The performance characteristics are required to determine if the method can be used for its intended purpose. The number of significant figures attached to the value of the characteristic generally indicates the reliability of these indices. They are generally limited by the repeatability standard deviation, sr. In most analytical work requiring calibration the best relative sr that can be achieved is about 1%. This is equivalent to the use of 2 significant figures. However, in order to avoid loss of "accuracy" in averaging operations, carry one additional figure with all reported values, i.e., use at most 3 significant figures in reporting. This statement, however, does not apply to recorded raw data, such as weighing or instrument readings, calibration, and standardization, which should utilize the full reading capacity of the measurement scales. This exception is limited by the measurement scale with the least reading capacity.

The purpose of the analysis determines which attributes are important and which may be less so.

3.1 Applicability (Scope)

A method must demonstrate acceptable recovery and repeatability with representative matrices and concentrations to which it is intended to be applied. For single materials, use at least three typical specimens, at least in duplicate, with different attributes (appearance, maturity, varieties, age). Repeat the analyses at least one day later. The means should not differ significantly and the repeatability should approximate those listed in *Section 3.4.2* for the appropriate concentration. If the method is intended to be applied to a single commodity, e.g., fruits, cereals, fats, use several representative items of the commodity with a range of expected analyte concentrations. If the method is intended to apply to "foods" in general, select representative items from the food triangle [Sullivan, D.M., & Carpenter, D.E. (1993) "Methods of Analysis for Nutrition Labeling," AOAC INTERNATIONAL, Gaithersburg, MD, pp 115–120]. In the case of residues, the matrices are generalized into categories such as "fatty foods" and "nonfatty foods" that require different preliminary treatments

to remove the bulk of the "inert" carrier. In all cases, select test materials that will fairly represent the range of composition and attributes that will be encountered in actual practice. Applicability may be inferred to products included within tested extremes but cannot be extrapolated to products outside the tested limits.

Similarly the range of expected concentrations should be tested in a number of typical matrices, spiking if necessary, to ensure that there is no interaction of analyte with matrix.

Semipermanent "house standards" for nutrients often can be prepared from a homogeneous breakfast cereal for polar analytes and from liquid monounsaturated oil like olive oil for nonpolar analytes for use as concurrent controls or for fortification.

The authority for the authenticity of botanical specimens and their source and the origin or history of the test materials must be given.

The determination of freedom from the effects of interfering materials is tested under selectivity, *Section 3.2*, and properties related to the range of quantification of the target analyte are tested under the reliability characteristics, *Section 3.4*.

3.2 Selectivity

The term selectivity is now generally preferred by IUPAC over specificity.

Selectivity is the degree to which the method can quantify the target analyte in the presence of other analytes, matrices, or other potentially interfering materials. This is usually achieved by isolation of the analyte through selective solvent extraction, chromatographic or other phase separations, or by application of analyte-specific techniques such as biochemical reactions (enzymes, antibodies) or instrumentation [nuclear magnetic resonance (NMR), infrared, or mass spectrometry (MS)].

Methods must be tested in the presence of accompanying analytes or matrices most likely to interfere. Matrix interference is usually eliminated by extraction procedures and the desired analyte is then separated from other extractives by chromatography or solid-phase extraction. Nevertheless, many methods for low-level analytes still require a matrix blank because of the presence of persistent, nonselective background.

The most useful separation technique is chromatography and the most important requirement is resolution of the desired peak from accompanying peaks. Resolution, R_s, is expressed as a function of both the absolute separation distance expressed as retention times (minutes) of the two peaks, t_1 and t_2, and the baseline widths, W_1 and W_2, of the analyte and nearest peak, also expressed in terms of times, as

$$R_s = 2 (t_2 - t_1) / (W_1 + W_2)$$

Baseline widths are measured by constructing tangents to the two sides of the peak band and measuring the distance between the intersection of these tangents with the baseline or at another convenient position such as half-height. A resolution of at least 1.5 is usually sought and one of 1.0 is the minimum usable separation. The U.S. Food and Drug Administration (FDA) suggests an R_s of at least 2 for all compounds accompanying active drug dosage forms, including hydrolytic, photolytic, and oxidative degradation products. In addition, the isolated analyte should show no evidence of other compounds when chromatographed on other systems consisting of different columns and solvents, or when examined by techniques utilized for specificity (infrared, NMR, or MS). These requirements were developed for synthetic drug substances, and must be relaxed for the families of compounds commonly encountered in foods and botanical specimens to a resolution of 1.5 from adjacent nontarget peaks.

If the product is mixed with other substances, the added substances must be tested to ensure that they do not contain any material that will interfere with the identification and determination of the analyte sought. If the active constituent is a mixture, the necessity for separation of the ingredients is a decision related to the complexity of the potential separation, the constancy of the relationship of the components, and the relative biological activity of the constituents.

3.3 Calibration

Modern instrumental methods depend upon the comparison of a signal from the unknown concentration of an analyte to that from a known concentration of the same or similar analyte. This requires the availability of a reference standard, *Section 2.2.2*. The simplest calibration procedure requires preparation of a series of standard solutions from the reference material, by dilution of a stock solution, covering a reasonable range of signal response from the instrument. Six to 8 points, approximately equally spaced over the concentration range of interest, performed in duplicate but measured at random (to avoid confusing nonlinearity with drift) is a suitable calibration pattern. Fit the calibration line (manually or numerous statistical and spreadsheet programs are available) and plot the residuals (the difference of the experimental points from the fitted line) as a function of concentration. An acceptable fit produces a random pattern of residuals with a 0 mean. For checking linearity, prepare the individual solutions by dilution from a common stock solution to avoid the random errors likely to be introduced from weighing small (mg) quantities for individual standards.

As long as the purity of the reference material is 95% or greater, as determined by evaluating secondary peaks or spots in gas, liquid, or thin-layer chromatography or other quantitative technique, the impurities contributes little to the final variance at micro- and ultramicro concentrations and may be neglected. (Recovery trials, however, require greater purity or correction for the impurities.) The identity of the material used as the reference material, however, is critical. Any suggestion of nonhomogenity such as multiple or distorted peaks or spots, insoluble residue, or appearance of new peaks on standing requires further investigation of the identity of the standard.

Similarly, certified volumetric glassware may also be used after initial verification of their stated capacity by weighing the indicated volume of water for flasks and the delivered volume for pipets and burets and converting the weight to the volume delivered.

Do not use serological pipets at less than 10% of their graduated capacity. Check the stability of the stock and initial diluted solutions, stored at room or lower temperatures, by repeating their measurements several days or weeks later. Prepare the most dilute solutions fresh as needed from more concentrated, stable solutions in most cases. Bring solutions stored at refrigerator or lower temperatures to room temperature before opening and using them.

Plot the signal response against the concentration. A linear response is desirable as it simplifies the calculations, but it is not necessary nor should it be regarded as a required performance characteristic. If the curve covers several orders of magnitude, weighted regression, easily handled by computer programs, may be useful. Responses from electrochemical and immunological methods are exponential functions, which often may be linearized by using logarithms. Some instruments perform signal-to-concentration calculations automatically using disclosed or undisclosed algorithms. If the method is not used routinely, several standards should accompany

the test runs. If the method is used routinely, the standard curve should be repeated daily or weekly, depending on its stability. Repeat the standard curve as frequently as necessary with those instruments where drift is a significant factor.

A high correlation coefficient (e.g., >0.99) is often recommended as evidence of goodness of fit. Such use of the correlation coefficient as a test for linearity is incorrect [Analytical Methods Committee, *Analyst* **113**, 1469–1471(1988); **119**, 2363(1994)]. Visual examination is usually sufficient to indicate linearity or nonlinarity, or use the residual test, *Section 3.3*.

If a single (parent or associated) compound is used as the reference material for a series of related compounds, give their relationship in structure and response factors.

Note that the calibration is performed directly with the analyte reference solutions. If these reference solutions are carried through the entire procedure, losses in various steps of the procedure cannot be explored but are automatically compensated for. Some procedures require correction of the final result for recovery. When this is necessary, use a certified reference material, a "house" standard, or analyte added to a blank matrix conducted through the entire method for this purpose. If several values are available from different runs, the average is usually the best estimate of recovery. Differences of calibration curves from day to day may be confused with matrix effects because they are often of the same magnitude.

3.3.1 External Standard Method

The most common calibration procedure utilizes a separately prepared calibration curve because of its simplicity. If there is a constant loss in the procedure, this is handled by a correction factor, as determined by conducting a known amount of analyte through the entire procedure. The calculation is based on the ratio of the response of equal amounts of the standard or reference compound to the test analyte. This correction procedure is time consuming and is used as a last resort since it only improves accuracy at the expense of precision. Alternatives are the internal standard procedure, blank matrix process, and the method of standard addition.

If the method is intended to cover a substantial range of concentrations, prepare the curve from a blank and five or seven approximately equally spaced concentration levels and repeat on a second day. Repeat occasionally as a check for drift. If an analyte is examined at substantially different concentration levels, such as pesticide residues and formulations, prepare separate calibration curves covering the appropriate range to avoid excessive dilutions. In such cases, take care to avoid cross contamination. However, if the analyte always occurs at or near a single level as in a pharmaceutical, a 2-point curve may be used to bracket the expected level, or even a single standard point, if the response over the range of interest is approximately linear. By substituting an analyte-free matrix preparation for the blank, as might be available from pesticide or veterinary drug residue studies or the excipients from a pharmaceutical, a calibration curve that automatically compensates for matrix interferences can be prepared.

3.3.2 Internal Standard Method

The internal standard method requires the addition of a known amount of a compound that is easily distinguished from the analyte but which exhibits similar chemical properties. The response ratio of the internal standard to a known amount of the reference standard of the analyte of interest is determined beforehand. An amount of internal standard similar to that expected for the analyte is added at an early stage of the method. This method is particularly useful for addition to the eluate from an HPLC separation when the fractions are held in an autosampler that is run overnight, where it compensates for any losses of solvent by evaporation. An internal standard is also frequently used in GLC residue methods where many analytes with similar properties are frequently encountered.

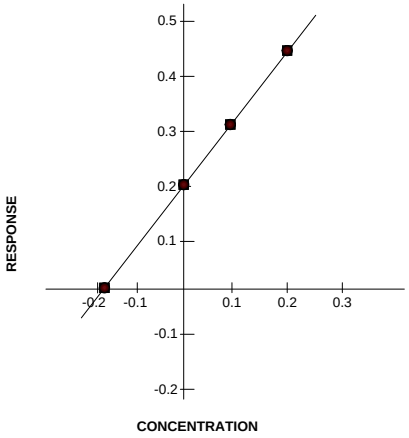


Figure 1

3.3.3 Standard Addition Method

When the matrix effect on an analyte is unknown or variable, the method of standard additions is useful. Make measurements on the isolated analyte solution and add a known amount of the standard analyte at the same level and at twice or three (or known fractions) times the original level. Plot the signal against the concentration with the initial unknown concentration set at 0. Extrapolate the line connecting the measured responses back to 0 response and read the concentration value off the (negative) *x*-axis. The main assumption is that the response is linear in the working region. This method is used most frequently with emission spectroscopy, electrochemistry, and radiolabeled isotopes in mass spectrometric methods.

See Figure 1 for example [from Rubinson, K.A. (1987) "Chemical Analysis," Little, Brown and Co., Boston, MA, p. 205].

Concn Cu added, μg	Instrument response
0.0	0.200
0.10	0.320
0.20	0.440
Concn Cu found by extrapolation to 0.00 response	(–)0.18

3.4 Reliability Characteristics

These are the statistical measures of how good the method is. Different organizations use different terms for the same concept. The important questions are:

- How close is the reported value to the true, reference, or accepted value?
- How close are repeated values to each other as determined in the same or different laboratories?
- What is the smallest amount or concentration that can be recognized or measured?

Recently accreditation organizations have been requesting the calculation of the parameter "Measurement Uncertainty" (MU). This is a term indicative of the reliability of the particular series of

measurements being reported. The standard uncertainty is equal to the standard deviation of the series of measurements of the analyte. The expanded uncertainty is two times the standard uncertainty and is expected to encompass about 95% of similar future measurements. If too few values are available in a measurement series to calculate a stable MU, the standard deviation obtained from the validation study within the laboratory, s_r, may be substituted, if it covered the same or similar analyte/matrix/concentration range. If a collaboratively studied method is being validated for use within a laboratory, the standard deviation among-laboratories, s_R, reported for the method from the study should be used to determine if the anticipated measurement uncertainty will be satisfactory for the intended purpose, assuming satisfactory repeatability as demonstrated by control charts or proficiency testing. In fact, the determination of the reliability characteristics in the validation study should not be undertaken until the developmental work demonstrates that the data are repeatable and in statistical control.

The Codex Alimentarius, an international body organized by the Food and Agricultural Organization (FAO) and the World Health Organization (WHO) of the United Nations (UN) to recommend international food standards to governments, suggests the following "Guidelines for the Assessment of the Competence of Testing Laboratories Involved in the Import and Export Control of Food" (FAO, Rome, Italy, CAC/GL 27-1997) for laboratories:

- Comply with the general competence criteria of ISO 17025
- Participate in proficiency testing schemes for food analysis
- Utilize validated methods
- Utilize internal quality control procedures

3.4.1 Accuracy

The term "accuracy" has been given so many meanings that it is better to use a more specific term. Ordinarily it means closeness of the test result to the "true" or accepted value. But the test result can be an individual value, the average of a set of values, or the average of many sets of values. Therefore, whenever the term is used, the number of values it represents and their relationship must always be stated, e.g., as an individual result, as the average of duplicates or n replicates, or as the average of a set of a number of trials. The difference of the reported value from the accepted value, whether it is an individual value, an average of a set of values, or the average of a number of averages, or an assigned value, is the bias under the reported conditions. The frequently used term for bias or "accuracy" when the average of a set of values is reported is "trueness."

The fraction or percentage of the analyte that is recovered when the test sample is conducted through the entire method is the recovery. The best reference materials for determining recovery are analyte-certified reference materials (CRMs) distributed by national metrological laboratories, but in most cases material certified by a commercial supplier must be accepted. Occasionally standards are available from a government agency, such as pesticides from the Environmental Protection Agency (EPA). They are rarely, if ever, available in the matrix of interest but rather as a solution in a convenient solvent with a stated concentration and uncertainty. Such reference materials must then be tested in the matrix of interest. Even rarer is an isotopically labeled analyte that can be easily followed by isotopic analytical techniques.

The available certified or commercial analyte standard, diluted if necessary, is added to typical analyte-free matrices at levels about 1x or 2x the expected concentration. Analyte-free matrices for residues are obtained from growers who certify that the chemical is not used in their cultivation, growth, or feeding and verified analytically. They may also be obtained from the residues of previously extracted materials or from test samples shown to be negative for the analyte.

If an analyte-free matrix is not available, the analyte standard is added to separate test portions and the recovery is calculated from the base determined by the method of addition, *Section 3.3.3*. Run the set of such controls with each set of test samples. If a sufficient number of batches are expected to be run (at least 20–30), the % recovery can be plotted against the run number as the basis for a control chart. Recovery also can be obtained as a byproduct of the precision determinations, *Sections 3.4.2* and *3.4.4*.

Acceptable recovery is a function of the concentration and the purpose of the analysis. Some acceptable recovery requirements for individual assays are as follows:

Concentration	Recovery limits, %
100%	98–101
10%	95–102
1%	92–105
0.1%	90–108
0.01%	85–110
10 μg/g (ppm)	80–115
1 μg/g	75–120
10 μg/kg (ppb)	70–125

The Codex Alimentarius "Residues of Veterinary Drugs in Foods" [2nd Ed., Vol. 3 (1993) Joint FAO/WHO Food Standards Program, FAO, Rome, Italy, p. 59] suggests the following limits for residues of veterinary drugs in foods:

Concentration, μg/kg	Acceptable range
≤1	50–120
≥1 < 10	60–120
≥10 < 100	70–110
≥100	80–110

These limits may be modified as needed in view of the variability of individual results or which set of regulatory requirements are referenced. (As a rough guide to typical performance, about 95% of normally distributed typical results in a single laboratory at 1 μg/g will fall within 80–120% of the mean.) In the case of the examination of the general USDA pesticide residue proficiency study, limits of 50–150% were applied; the USFDA acceptability criterion for recovery of drug residues at the 10 ppb level is 70–120%. Generally, however, recoveries less than 60–70% should be subject to investigations leading to improvement and average recoveries greater than 110% suggest the need for better separations. Most important, recoveries greater than 100% must not be discarded as impossible. They are the expected positive side from a typical distribution of analytical results from analytes present at or near 100% that are balanced by equivalent results on the negative side of the mean.

If an extraction of active ingredient from a matrix with a solvent is used, test extraction efficiency by reextracting the (air-dried) residue and determining the active ingredient(s) in the residue by the method.

The number of units to be used to establish bias is arbitrary, but the general rule is the more independent "accuracy" trials, the better. The improvement, as measured by the width of the confidence interval for the mean, follows the square root of the number of trials. Once past 8–10 values, improvement comes slowly. To fully contribute, the values must be conducted independently, i.e., nonsimultaneously, throwing in as many environmental or spontaneous differences as possible, such as different analysts, instruments, sources of reagents, time of day,

temperature, barometric pressure, humidity, power supply voltage, etc. Each value also contributes to the within-laboratory precision as well. A reasonable compromise is to obtain 10 values from a reference material, a spiked matrix, or by the method of standard addition scattered over several days or in different runs as the basis for checking bias or recovery. By performing replicates, precision is obtained simultaneously. Precision obtained in such a manner is often termed "intermediate precision" because its value is between within-laboratory and among-laboratory precision. When reported, the conditions that were held constant and those that were varied must be reported as well.

Note that the series of determinations conducted for the method of addition are not independent because they are probably prepared from the same standard calibration solution, same pipets, and are usually conducted almost simultaneously. This is satisfactory for their intended purpose of providing an interrelated function, but it is not satisfactory for a precision function estimation intended for future use.

Related to recovery is the matter of reporting the mean corrected or not corrected for recovery. Unless specifically stated in the method to correct or not, this question is usually considered a "policy" matter and is settled administratively outside the laboratory by a regulatory pronouncement, informal or formal agreement, or by contract. If for some reason a value closest to theory is needed, correction is usually applied. If a limit or tolerance has been established on the basis of analytical work with the same method correlated with "no effect" levels, no correction should be applied because it has already been used in setting the specification. Corrections improve "accuracy" at the expense of impairing precision because the variability of both the determination and the recovery are involved.

When it is impossible to obtain an analyte-free matrix to serve as a base for reporting recovery, two ways of calculating recovery must be distinguished: (*1*) Total recovery based on recovery of the native plus added analyte, and (*2*) marginal recovery based only on the added analyte (the native analyte is subtracted from both the numerator and denominator). Usually total recovery is used unless the native analyte is present in amounts greater than about 10% of the amount added, in which case use the method of addition, *Section 3.3.3*.

When the same analytical method is used to determine both the concentration of the fortified, C_f, and unfortified, C_u, test samples, the % recovery is calculated as

$$\text{Recovery, \%} = (C_f - C_u) \times 100/C_a$$

where C_a is the calculated (not analyzed) concentration of analyte added to the test sample. The concentration of added analyte should be no less that the concentration initially present and the response of the fortified test sample must not exceed the highest point of the calibration curve. Both fortified and unfortified test samples must be treated identically in the analysis.

3.4.2 Repeatability Precision (s_r, RSD_r)

Repeatability refers to the degree of agreement of results when conditions are maintained as constant as possible with the same analyst, reagents, equipment, and instruments performed within a short period of time. It usually refers to the standard deviation of simultaneous duplicates or replicates, s_r. It is the best precision that will be exhibited by a laboratory but it is not necessarily the laboratory's typical precision. Theoretically the individual determinations should be independent but this condition is practically impossible to maintain when determinations are conducted simultaneously and therefore this requirement is generally ignored.

To obtain a more representative value for the repeatability precision perform the simultaneous replicates at different times (but the same day), on different matrices, at different concentrations. Calculate the standard deviation of repeatability from at least five pairs of values obtained from at least one pair of replicates analyzed with each batch of analyses for each pertinent concentration level that differs by approximately an order of magnitude and conducted at different times. The object is to obtain representative values, not the "best value," for how closely replicates will check each other in routine performance of the method. Therefore these sets of replicate analyses should be conducted at least in separate runs and preferably on different days. The repeatability standard deviation varies with concentration, C expressed as a mass fraction. Acceptable values approximate the values in the following table or calculated by the formula:

$$RSD_r, \% = 2C^{-0.15}$$

unless there are reasons for using tighter requirements.

Concentration	Repeatability (RSD_r), %
100%	1
10%	1.5
1%	2
0.1%	3
0.01%	4
10 μg/g (ppm)	6
1 μg/g	8
10 μg/kg (ppb)	15

Acceptable values for repeatability are between ½ and 2 times the calculated values. Alternatively a ratio can be calculated of the found value for RSD_r to that calculated from the formula designated as $HorRat_r$. Acceptable values for this ratio are typically 0.5 to 2:

$$HorRat_r = RSD_r \text{ (found, \%)}/RSD_r \text{ (calculated, \%)}$$

The term "repeatability" is applied to parameters calculated from simultaneous replicates and this term representing minimum variability is equated to the "within-laboratory" parameter (standard deviation, variance, coefficient of variation, relative standard deviation) of the precision model equation. It should be distinguished from a somewhat larger within-laboratory variability that would be induced by non-simultaneous replicates conducted in the same laboratory on identical test samples on different days, by different analysts, with different instruments and calibration curves, and with different sources of reagents, solvents, and columns. When such an "intermediate" within-laboratory precision (standard deviation, variance, coefficient of variation, relative standard deviation) is used, a statement of the conditions that were not constant must accompany it. These within-laboratory conditions have also been called within-laboratory reproducibility, an obvious misnomer.

3.4.3 Measurement Uncertainty

Accreditation organizations have been requesting laboratories to have a parameter designated as "measurement uncertainty" associated with methods that the laboratory utilizes. The official metrological definition of measurement uncertainty is "a parameter

associated with the result of a measurement that characterizes the dispersion of values that could reasonably be attributed to the measurand." A note indicates, "the parameter may be, for example, a standard deviation (or a given multiple of it), or the width of a confidence interval."

Of particular pertinence is the fact that the parameter applies to a measurement and not to a method (*see Section 3.4*). Therefore "standard" measurement uncertainty is the standard deviation or relative standard deviation from a series of simultaneous measurements. "Expanded" uncertainty is typically twice the standard uncertainty and is considered to encompass approximately 95% of future measurements. This is the value customarily used in determining if the method is satisfactory for its intended purpose although it is only an approximation because theoretically it applies to the unknown "true" concentration.

Since the laboratory wants to know beforehand if the method will be satisfactory for the intended purpose, it must use the parameters gathered in the validation exercises for this purpose, substituting the measurement values for the method values after the fact. As pointed out by M. Thompson [*Analyst* **125**, 2020–2025 (2000); *see Inside Lab. Mgmt*. **5**(2), 5(2001)], a ladder of errors exist for this purpose.

- Duplicate error (a pair of tests conducted simultaneously)
- Replicate or run error (a series of tests conducted in the same group)
- Within-laboratory error (all tests conducted by a laboratory)
- Between-laboratory error (all tests by all laboratories)

As we go down the series, the possibility of more errors being included is increased until a maximum is reached with the all inclusive reproducibility parameters. Thompson estimates the relative magnitude of the contribution of the primary sources of error as follows

Level of variation	Separate	Cumulative
Repeatability	1.0	1.0
Runs	0.8	1.3
Laboratories	1.0	1.6
Methods	1.5	2.2

Ordinarily only one method exists or is being validated so we can ignore the last line. Equating duplicates to replicability, runs to within-laboratory repeatability, and laboratories to among-laboratories reproducibility, Thompson points out that the three sources of error are roughly equal and not much improvement in uncertainty would result from improvement in any of these sources. In any case, the last column gives an approximate relative relationship of using the standard deviation at any point of the ladder as the basis for the uncertainty estimate prior to the actual analytical measurements.

In the discussion of uncertainty it must be noted that bias as measured by recovery is not a component of uncertainty. Bias (a constant) should be removed by subtraction before calculating standard deviations. Differences in bias as exhibited by individual laboratories become a component of uncertainty through the among-laboratory reproducibility. The magnitude of the uncertainty depends on how it is used—comparisons within a laboratory, with other laboratories, and even with other methods. Each component adds uncertainty. Furthermore, uncertainty stops at the laboratory's edge. If only a single laboratory sample has been submitted and analyzed, there is no basis for estimating sampling uncertainty. Multiple independent samples are required for this purpose.

3.4.4 Reproducibility Precision (s_R, RSD_R)

Reproducibility precision refers to the degree of agreement of results when operating conditions are as different as possible. It usually refers to the standard deviation (s_R) or the relative standard deviation (RSD_R) of results on the same test samples by different laboratories and therefore is often referred to as "between-laboratory precision" or the more grammatically correct "among-laboratory precision." It is expected to involve different instruments, different analysts, different days, and different laboratory environments and therefore it should reflect the maximum expected precision exhibited by a method. Theoretically it consists of two terms: the repeatability precision (within-laboratory precision, s_r) and the "true" between-laboratory precision, s_L. The "true" between-laboratory precision, sL, is actually the pooled constant bias of each individual laboratory, which when examined as a group is treated as a random variable. The between-laboratory precision too is a function of concentration and is approximated by the Horwitz equation, $s_R = 0.02C^{0.85}$. The AOAC/IUPAC protocol for interlaboratory studies requires the use of a minimum of eight laboratories examining at least five materials to obtain a reasonable estimate of this variability parameter, which has been shown to be more or less independent of analyte, method, and matrix.

By definition s_R does not enter into single-laboratory validation. However, as soon as a second (or more) laboratory considers the data, the first question that arises involves reanalysis by that second laboratory: "If I had to examine this or similar materials, what would I get?" As a first approximation, in order to answer the fundamental question of validation—fit for the intended purpose—assume that the recovery and limit of determination are of the same magnitude as the initial effort. But the variability, now involving more than one laboratory, should be doubled because variance, which is the square of differences, is involved, which magnifies the effect of this parameter. Therefore we have to anticipate what another laboratory would obtain if it had to validate the same method. If the second laboratory on the basis of the doubled variance concludes the method is not suitable for its intended purpose, it has saved itself the effort of revalidating the method.

In the absence of such an interlaboratory study, the interlaboratory precision may be estimated from the concentration as indicated in the following table or by the formula (unless there are reasons for using tighter requirements):

$$RSD_R = 2C^{-0.15}$$

or

$$S_R = 0.02C^{0.85}$$

Concentration, C	Reproducibility (RSD_R), %
100%	2
10%	3
1%	4
0.1%	6
0.01%	8
10 μg/g (ppm)	11
1 μg/g	16
10 μg/kg (ppb)	32

Acceptable values for reproducibility are between ½ and 2 times the calculated values. Alternatively a ratio can be calculated

of the found value for RSD_R to that calculated from the formula designated as $HorRat_R$. Acceptable values for this ratio are typically 0.5 to 2:

$$HorRat_R = RSD_R \text{ (found, \%)}/RSD_R \text{ (calculated, \%)}$$

As stated by Thompson and Lowthian ("The Horwitz Function Revisited," (1997) *J. AOAC Int.* **80**, 676–679), "Indeed, a precision falling within this 'Horwitz Band' is now regarded as a criterion for a successful collaborative trial."

The typical limits for HorRat values may not apply to indefinite analytes (enzymes, polymers), physical properties, or to the results from empirical methods expressed in arbitrary units. Better than expected results are often reported at both the high (>10%) and low (<E-8) ends of the concentration scale. Better than predicted results can also be attained if extraordinary effort or resources are invested in education and training of analysts and in quality control.

3.4.5 Intermediate Precision

The precision determined from replicate determinations conducted within a single laboratory not simultaneously, i.e., on different days, with different calibration curves, with different instruments, by different analysts, etc. is called intermediate precision. It lies between the within- and among-laboratories precision, depending on the conditions that are varied. If the analysis will be conducted by different analysts, on different days, on different instruments, conduct at least five sets of replicate analyses on the same test materials under these different conditions for each concentration level that differs by approximately an order of magnitude.

3.4.6 Limit of Determination

The limit of determination is a very simple concept: It is the smallest amount or concentration of an analyte that can be estimated with acceptable reliability. But this statement contains an inherent contradiction: the smaller the amount of analyte measured, the greater the unreliability of the estimate. As we go down the concentration scale, the standard deviation increases to the point where a substantial fraction of values of the distribution of results overlaps 0 and false negatives appear. Therefore the definition of the limit comes down to a question of what fraction of values are we willing to tolerate as false negatives.

Thompson and Lowthian (loc. cit.) consider the point defined by $RSD_R = 33\%$ as the upper bound for useful data, derived from the fact that $3RSD_R$ should contain 100% of the data from a normal distribution. This is equivalent to a concentration of about 8×10^{-9} (as a mass fraction) or 8 ng/g (ppb). Below this level false negatives appear and the data goes "out of control." From the formula, this value is also equivalent to an $RSD_r \approx 20\%$. The penalty for operating below the equivalent concentration level is the generation of false negative values. Such signals are generally accepted as negative and are not repeated.

An alternative definition of the limit of detection and limit of determination is based upon the variability of the blank. The blank value, x_{Bl}, plus 3 times the standard deviation of the blank ($x_{Bl} + 3s_{Bl}$) is taken as the detection limit and the blank value plus 10 times the standard deviation of the blank ($x_{Bl} + 10s_{Bl}$) is taken as the determination limit. The problem with this approach is that the blank is often difficult to measure or is highly variable. Furthermore, the value determined in this manner is independent of the analyte. If blank values are accumulated over a period of time, the average is likely to be fairly representative as a basis for the limits and will probably provide a value of the same magnitude as that derived from the relative standard deviation formulae.

The detection limit is only useful for control of undesirable impurities that are specified as "not more than" a specified low level and for low-level contaminants. Useful ingredients must be present at high enough concentrations to be functional. The specification level must be set high enough in the working range that acceptable materials do not produce more than 5% false-positive values, the default statistical acceptance level. Limits are often at the mercy of instrument performance, which can be checked by use of pure standard compounds. Limits of detection and determination are unnecessary for composition specifications although the statistical problem of whether or not a limit is violated is the same near zero as it is at a finite value.

Blank values must be monitored continuously as a control of reagents, cleaning of glassware, and instrument operation. The necessity for a matrix blank would be characteristic of the matrix. Abrupt changes require investigation of the source and correction. Taylor [J.K. Taylor (1987) "Quality Assurance of Chemical Measurements," Lewis Publishers, Chelsea, MI, p. 127] provides two empirical rules for applying a correction in trace analysis: (*1*) The blank should be no more than 10% of the "limit of error of the measurement", and (*2*) it should not exceed the concentration level.

3.4.7 Reporting Low-Level Values

Although on an absolute scale low level values are miniscule, they become important in three situations:

(*1*) When legislation or specifications decrees the absence of an analyte (zero tolerance situation).

(*2*) When very low regulatory or guideline limits have been established in a region of high uncertainty (e.g., a tolerance of 0.005 μg/kg aflatoxin M_1 in milk).

(*3*) When dietary intakes of low-level nutrients or contaminants must be determined to permit establishment of minimum recommended levels for nutrients and maximum limits for contaminants.

Analytical work in such situations not only strains the limits of instrumentation but also the ability of the analyst to interpret and report the findings. Consider a blank that is truly 0 and that the 10% point of the calibration curve corresponds to a concentration of 1 μg/kg (E-9). By the Horwitz formula this leads to an expected RSD_r in a single laboratory of about 23%. If we assume a normal distribution and we are willing to be wrong 5% of the time, what concentration levels would be expected to appear? From 2-tail normal distribution tables (the errant value could appear at either end), 2.5% of the values will be below 0.72 μg/kg and 2.5% will be above 1.6 μg/kg. Note the asymmetry of the potential results, from 0.7 to 1.6 μg/kg for a nominal 1.0 μg/kg value from the nature of the multiplicative scale when the RSD is relatively large.

But what does the distribution look like at zero? Mathematically it is intractable because it collapses to zero. Practically, we can assume the distribution looks like the previous one but this time we will assume it is symmetrical to avoid complications. The point to be made will be the same. For a distribution to have a mean equal to 0, it must have negative as well as positive values. But negative concentration values per se are forbidden but here they are merely an artifact of transforming measured signals. Negative signals are typical in electromotive force and absorbance measurements.

Analysts have an aversion to reporting a zero concentration value because of the possibility that the analyte might be present, but below the detection limit. Likewise, analysts avoid reporting

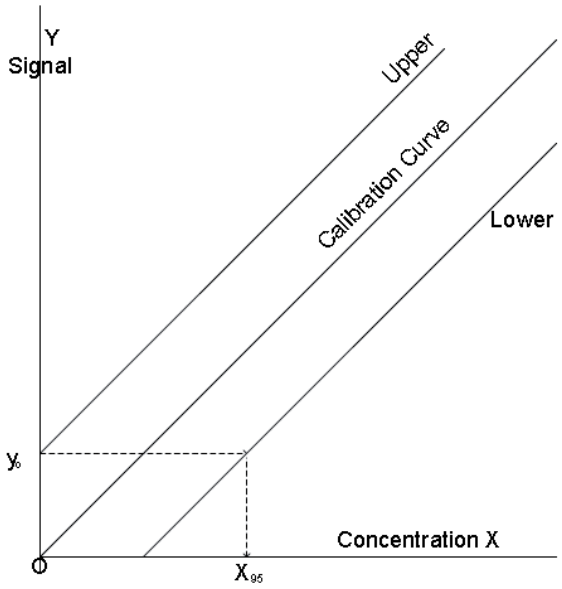


Figure 2. The statistical situation at the zero concentration level: A signal as high as y_0 could be measured at a 0 concentration, which corresponds to a "true" concentration value as high as x_{95}, but with only a 5% probability.

negative values as physical impossibilities although they are required by arithmetic averaging of random fluctuations to attain a real zero. Analysts avoid the issue by linguistic subterfuges such as "less than the detection limit" or by substituting an arbitrary fractional value such as one half the detection limit. Statisticians must discard such values as useless and consequently much effort is simply wasted by such reports.

Therefore the recommendation for handling low level values for validation purposes is to report whatever value is returned by converting the recorded instrument reading to a concentration through the calibration chart: positive, negative, or zero and rely on the power of averaging to produce the best estimate. As stated by the (UK) Analytical Methods Committee (*Anal. Tech. Brief No. 5*, April 2001), "analytical results are not concentrations but error-prone estimates of concentrations."

Such advice is impractical for reporting to a nontechnical or even a technical reviewer unfamiliar with the statistical problem of reporting results near zero. In such cases, the simplest solution is to report "zero" or "none found" for all signal values within the region of (blank value + 3 x (standard deviation of the blank signal)). This can be supplemented by a statement that the variability of results in the region of zero is such that it would permit as much as *x* µg/kg to be present with not more than a 5% probability, where *x* is roughly 5. If the laboratory can calculate the confidence interval of the calibration curve, a better estimate is obtained by drawing a line parallel to the *x*-axis from the *y* (signal) value where the upper confidence line intersects the *y*-axis (y_0) until it intersects the lower confidence line and reading the *x* (concentration) value (x_{95}) of the line parallel to the *y*-axis where it intersects the *x*-axis (*see* Figure 2). This curve can be used to supply a statement that any signal less than y_0 can be reported as "zero" or "none found" with only a 5% chance of being wrong.

3.4.8 Dichotomous Reporting (Qualitative Analysis)

In an effort to bypass the laborious effort to develop and validate a method of analysis, a request is often made to obtain a test that will merely verify the presence or absence of an analyte. Such a request assumes correctly that it is simpler to divide a continuum of measurements of a property into two parts than into more than two parts. This concept assigns all values on one side of the division as acceptable, positive, or present and all values on the other side as unacceptable, absent, or negative. Even assuming that it is easy to set a dividing value through an external specification, tolerance, or limit-setting procedure, we cannot escape the statistical problem of interpretation of a measured value because of the accompanying distribution or halo of uncertainty.

This problem was discussed many years ago in connection with the interpretation of very simple spot tests by Feigl, the developer of this technique [Feigl, F. (1943) "Laboratory Manual of Spot Tests," Academic Press, New York, NY]. "If the sensitivity of a spot reaction is checked by progressively diluting a given standard solution, and then at each dilution, one drop is taken for the test, different observers will practically never agree absolutely in their determinations of the identification limit, even though the same experimental conditions have been closely maintained by all. Almost always there will be a certain range of variation." (p. 4)

We now understand the reason for the "range of variation." It arises from the statistical distribution of any physical measurement characterized by a location parameter (mean) and a distribution parameter (standard deviation). Any single observation removed from the distribution at the dividing value could have been anywhere within the envelope of that distribution. Half of the observations will be above and half below even though the "true value" of the property is a fixed number. The property may be fixed, but the measurements are variable.

A qualitative test has been defined in terms of indicating if an analyte is present or absent, above or below a limit value, and as a test with "poorer" precision than a quantitative method. But all of these definitions degenerate into the single test of whether a measured value is significantly different (in a statistical sense) from a fixed value.

Consequently when a test is used in a qualitative manner, any anticipated gain in the number of test samples examined at the expense of reliability, is illusionary. The test is fundamentally no different from determining if a found value is above or below a quantitative specification value. When the concentration drops into a region of high measurement variability the signal degenerates from real measurements into false positives for the blanks and false negatives for the measurements.

Nevertheless, the Codex Alimentarius "Residues of Veterinary Drugs in Foods" [Vol. 3, 2nd Ed. (1993) Joint FAO/WHO Food Standards Program, FAO, Rome, Italy, pp 55–59] recognizes such methods as a Level III method to determine the presence or absence of a compound "at some designated level of interest." It anticipates that such methods involve microbiological or immunological principles and they "should produce less than 5% false negatives and less than 10% false positives when analysis is performed on the test sample." It is doubtful if the statistical properties (e.g., power) of this recommendation have been examined and if such requirements are achievable with a reasonable number of examinations. A rough calculation indicates that to achieve the required specification more than 200 independent tests on the same test sample would have to be made, a requirement that would probably exhaust the analytical sample before a dozen tests were made.

3.5 Controls

3.5.1 Control Charts

Control charts are only useful for large volume or continuous work. They require starting with at least 20–30 values to calculate a mean and a standard deviation, which form the basis for control values equivalent to the mean ± 2 s_r (warning limits) and the mean ± 3 s_r (rejection limits). At least replicate test portions of a stable house reference material and a blank are run with every batch of multiple test samples and the mean and standard deviations (or range of replicates) of the controls and blank are plotted separately. The analytical process is "in control" if not more than 5% of the values fall in the warning zone. Any value falling above the rejection limit or two consecutive values in the warning region requires investigation and corrective action.

3.5.2 Injection Controls

A limit of 1 or 2% is often placed on the range of values of the peak heights or areas or instrument response of repeated injections of the final isolated analyte solution. Such controls are good for checking stability of the instrument during the time of checking but give no information as to the suitability of the isolation part of the method. Such a limit is sometimes erroneously quoted as a relative standard deviation when range is meant.

3.5.3 Duplicate Controls

Chemists will frequently perform their analyses in duplicate in the mistaken belief that if duplicates check, the analysis must have been conducted satisfactorily. ISO methods often require that the determinations be performed in duplicate. Simultaneous replicates are not independent—they are expected to check because the conditions are identical. The test portions are weighed out using the same weights, aliquots are taken with the same pipets, the same reagents are used, operations are performed within the same time frame, instruments are operated with the same parameters, and the same operations are performed identically. Under such restraints, duplicates that do not check would be considered as outliers. Nevertheless, the parameter calculated from duplicates within a laboratory is frequently quoted as the repeatability limit, r, as equal to $2*\sqrt{2}*s_r$ and is expected to encompass 95% of future analyses conducted similarly. The corresponding parameter comparing two values in different laboratories is the reproducibility limit, $R = 2*\sqrt{2}*s_R$. This parameter is expected to reflect more independent operations. Note the considerable difference between the standard deviations, s_r and s_R, an average-type parameter, and the repeatability and reproducibility limits, r and R, which are 2.8 times larger. If duplicates do not check within the r value, look for a problem—methodological, laboratory, or sample in origin. Note that these limits ($2*\sqrt{2} = 2.8$) are very close to the limits used for rejection in control charts $3*s_r$. Therefore they are most useful for large volume routine work rather than for validation of methods. Note the considerable difference between the standard deviations, s_r and s_R, an average-type parameter, and the repeatability and reproducibility limits, r and R, which are 2.8 times larger.

3.6 Confirmation of Analyte

Because of the existence of numerous chemical compounds, some of which have chemical properties very close to analytes of interest, particularly in chromatographic separations, but different biological, clinical, or toxicological properties, regulatory decisions require that the identity of the analyte of interest be confirmed by an independent procedure. This confirmation of chemical identity is in addition to a quantitative "check analysis," often performed independently by a second analyst to confirm that the quantity of analyte found in both analyses exceeds the action limit.

Confirmation provides unequivocal evidence that the chemical structure of the analyte of interest is the same as that identified in the regulation. The most specific method for this purpose is mass spectrometry following a chromatographic separation with a full mass scan or identification of three or four fragments that are characteristic of the analyte sought or the use of multiple mass spectrometric (MS^n) examination. Characteristic bands in the infrared can also serve for identification but this technique usually requires considerably more isolated analyte than is available from chromatographic separations unless special examination techniques are utilized. Visible and ultraviolet spectra are too subject to interferences to be useful, although characteristic peaks can suggest structural characteristics.

Other techniques that can be used for identification, particularly in combination, in approximate order of specificity, include:

(*1*) Co-chromatography, where the analyte, when mixed with a standard and then chromatographed by HPLC, GLC, or TLC, exhibits a single entity, a peak or spot with enhanced intensity.

(*2*) Characteristic fluorescence (absorption and emission) of the native compound or derivatives.

(*3*) Identical chromatographic and spectral properties after isolation from columns of different polarities or with different solvents.

Identical full-scan visible or ultra-violet spectra, with matching peak(s).

Furthermore, no additional peaks should appear when chromatographic conditions are changed, e.g., different solvents, columns, gradients, temperature, etc.

3.7 Stability of the Analyte

The product should be held under typical or exaggerated storage conditions and the active ingredient(s) assayed periodically for a period of time judged to reasonably exceed the shelf life of the product. In addition, the appearance of new analytes from deterioration should be explored, most easily by a fingerprinting technique, *Section 2.1*.

4 Report (as applicable)

4.1 Title

- Single-Laboratory Validation of the Determination of [Analyte] in [Matrix] by [Nature of Determination]
- Author, Affiliation
- Other Participants

4.2 Applicability (Scope)

- Analytes (common and chemical name; CAS registry number or Merck index number)
- Matrices used
- In presence of
- In absence of
- Safety statements applicable to product

4.3 Principle

- Preparation of test portion
- Extraction
- Purification

- Separation
- Measurement
- Alternatives
- Interferences

4.4 Reagents

(Reagents usually present in a laboratory need not be listed.)

- Reference standards, identity, source, purity
- Calibration standard solutions, preparation, storage, stability
- Solvents (special requirements)
- Buffers
- Others

4.5 Apparatus

(Equipment usually present in a laboratory need not be listed; provide source, Web address, and catalog numbers of special items.)

- Chromatographic equipment (operating conditions; system suitability conditions; expected retention times, separation times, peak or area relations)
- Temperature-controlled equipment
- Separation equipment (centrifuges, filters)
- Measurement instruments

4.6 Calibration

- Range, number and distribution of standards, replication, stability

4.7. Procedure

- List all steps of method, including any preparation of the test sample.
- Critical points
- Stopping points

4.8 Calculations

- Formulae, symbols, significant figures

4.9 Controls

4.10 Results of Validation

4.10.1 Identification Data

- Analytes measured and properties utilized (matrices tested; reference standard, source, identity, purity)

4.10.2 Performance Data

- Recovery of control material
- Repeatability (by replication of entire procedure on same test sample)
- Limit of determination]concentration where RSD_r = 20% or (blank + 10 * s_{blank})]
- Expanded measurement uncertainty $2*s_r$

4.10.3 Low-Level Data

Report instrument reading converted to a concentration through the calibration curve: positive, negative, or zero. Do not equate to 0, do not truncate data, or report "less than."

Interpretation: Concentrations less than 5 µg/kg may be reported as "zero" or "less than 5 µg/kg" with a 95% probability (5% chance of being incorrect).

4.10.4 Stability Data

ANNEX A
Abbreviations and Symbols Used

CAS	Chemical Abstracts Service (Registry Number)
CRM	Certified Reference Material
FDA	U.S. Food and Drug Administration
EPA	U.S. Environmental Protection Agency
GLC	Gas-liquid chromatography
HPLC	High-performance liquid chromatography
i	(as a subscript) Intermediate in precision terms
ISO	International Organization for Standardization
MU	Measurement Uncertainty
MS	Mass Spectrometry
MS^n	Multiple mass spectrometry
NMR	Nuclear magnetic resonance
r, R	Repeatability, reproducibility limits: The value less than or equal to the absolute difference between two test results obtained under repeatability (reproducibility) conditions is expected to be with a probability of 95% = $2*\sqrt{2}*s_r(s_R)$
RSD_r	Repeatability relative standard deviation = $s_r \times 100$
RSD_R	Reproducibility relative standard deviation = $s_R \times 100$
s_r	Repeatability standard deviation (within-laboratories)
s_R	Reproducibility standard deviation (among-laboratories)
$\bar{x}$	Mean, average

ANNEX B
Example of a Ruggedness Trial

Choose seven factors that may affect the outcome of the extraction and assign reasonable high and low values to them as follows:

Factor	High value	Low value
Weight of test portion	A = 1.00 g	a = 0.50 g
Extraction temperature	B = 30°	b = 20°
Volume of solvent	C = 100 mL	c = 50 mL
Solvent	D = Alcohol	d = Ethyl acetate
Extraction time	E = 60 min	e = 30 min
Stirring	F = Magnetically	f = Swirl 10 min intervals
Irradiation	G = Light	g = Dark

Conduct eight runs (a single analysis that reflects a specified set of factor levels) utilizing the specific combinations of high and low values for the factors as follows, and record the result obtained for each combination. (It is essential that the factors be combined exactly as specified or erroneous conclusions will be drawn.)

Run No.	Factor combinations	Measurement obtained
1	A B C D E F G	x1
2	A B c D e f g	x2
3	A b C d E f g	x3
4	A b c d e F G	x4
5	a B C d e F g	x5
6	a B c d E f G	x6
7	a b C D e f G	x7
8	a b c D E F g	x8

To obtain the effect of each of the factors, set up the differences of the measurements containing the subgroups of the capital letters and the small letters from column 2 thus:

Effect of A and a

$$[(x1 + x2 + x3 + x4)/4] - [(x5 + x6 + x7 + x8)/4] = J$$

$$4A/4 - 4a/4 = J$$

Note that the effect of each level of each chosen factor is the average of four values and that the effects of the 7 other factors

cancel out. (The Youden ruggedness trial or fractional factorial experiment was designed for this outcome.) Similarly,

Effect of B and b
$$[(x1 + x2 + x5 + x6)/4] – [(x3 + x4 + x7 + x8)/4] = K$$
$$4B/4 – 4b/4 = K$$

Effect of C and c
$$[(x1 + x3 + x5 + x7)/4] – [(x2 + x4 + x6 + x8)/4] = L$$
$$4C/4 – 4c/4 = L$$

Effect of D and d
$$[(x1 + x2 + x7 + x8)/4] – [(x3 + x4 + x5 + x6)/4] = M$$
$$4D/4 – 4d/4 = M$$

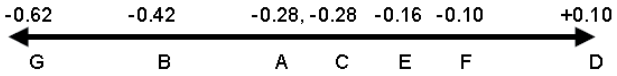
Effect of E and e
$$[(x1 + x3 + x6 + x8)/4] – [(x2 + x4 + x5 + x7)/4] = N$$
$$4E/4 – 4e/4 = N$$

Effect of F and f
$$[(x1 + x4 + x5 + x8)/4] – [(x2 + x3 + x6 + x7)/4] = O$$
$$4F/4 – 4f/4 = O$$

Effect of G and g
$$[(x1 + x4 + x6 + x7)/4] – [(x2 + x3 + x5 + x8)/4] = P$$
$$4G/4 – 4g/4 = P$$

Perform the eight determinations or runs carefully using the assigned factor level combinations and tabulate the values found. Then unscramble the 7 factors and obtain the effect of the assigned factor as the last number. It is important to use the combination of subscripts as assigned for proper interpretation.

Expt.	Found, %	Factors
x1	1.03	J (A) = 4A/4 – 4a/4 = 4.86 – 5.14 = –0.28
x2	1.32	K (B) = 4B/4 – 4b/4 = 4.79 – 5.21 = –0.42
x3	1.29	L (C) = 4C/4 – 4c/4 = 4.86 – 5.14 = –0.28
x4	1.22	M (D) = 4D/4 – 4d/4 = 5.05 – 4.95 = +0.10
x5	1.27	N (E) = 4E/4 – 4e/4 = 4.92 – 5.08 = –0.16
x6	1.17	O (F) = 4F/4 – 4f/4 = 4.95 – 5.05 = –0.10
x7	1.27	P (G) = 4G/4 – 4g/4 = 4.69 – 5.31 = –0.62
x8	1.43	



These values are plotted on a line. In this case they are more or less uniformly scattered along the line, but some attention should be paid to the extremes. Factor D, the highest positive value represents a difference in solvent, as expected, and this factor has to be investigated further to determine if the high values represents impurities or additional active ingredient. The extreme value of factor G suggests that the extraction should be conducted in the dark. As discussed by Youden, considerably more information can be obtained by utilizing several different materials and several independent replications in different laboratories, so as to obtain an estimate of the standard deviation to be expected between laboratories. Although the ruggedness trial is primarily a method development technique, validation of the application of a method to different matrices and related analytes can be explored simultaneously by this procedure.

Comments not used (may be added later):

3.3 Calibration: Run standards from low to high to compensate for any carryover. [Run in random order to compensate for drift is more important than allowing for carryover which should not occur.]

Independently made standards results in considerable random error in the calibration curve and is in fact the major source of random error in spectrophotometry. [Therefore a common stock solution is the preferred way of preparing the individual standards.]

Version 54 contains revisions as a result of comments from levanseler@nsf.org and McClure. Outline:

I. Types and benefits of each method validation study without reproducibility

II. Preparing for a Single-Laboratory Method Validation Study

III. Review of Performance Characteristics of a Method

IV. Errors

V. Calibration and Types

VI. Bias and Precision Estimations (no reference standard; no reproducibility)

VII. Detection and Quantification Limits

VIII. Ruggedness

PART II
AOAC Guidelines for Validation of Botanical Identification Methods

Contents

This document provides technical protocol guidelines for the AOAC validation of botanical identification methods and/or procedures, and covers terms and their definitions associated with the *Performance Tested Methods*SM and *Official Methods of Analysis*SM programs.

The guidelines working group consisted of James Harnly (Chair, USDA, ARS), Wendy Applequist (Missouri Botanical Garden), Paula Brown (British Columbia Institute of Technology), Steven Caspar (FDA/CFSAN), Peter Harrington (Ohio University), Danica Harbaugh-Reynaud (AuthenTechnologies, LLC), Norma Hill (Alcohol and Tobacco Tax and Trade Bureau Compliance Laboratory), Robert LaBudde (Least Cost Formulations and Old Dominion University), James Neal-Kababick (Flora Research Laboratories), Mark Roman (Tampa Bay Analytical Research), Shauna Roman (Schiff Nutrition International, Inc.), Darryl Sullivan (Covance Laboratories), Barry Titlow (Compound Solutions), and Paul Wehling (General Mills/Medallion Laboratories).

The guidelines were approved by the AOAC Official Methods Board on October 13, 2011.

This work was funded by the National Institutes of Health, Office of Dietary Supplements.

Reference: *J. AOAC Int*. **95**, 268–272(2012); DOI: 10.5740/jaoacint.11-447

1 Scope

The purpose of this document is to provide comprehensive technical guidance for conducting AOAC INTERNATIONAL (AOAC) validation studies for botanical identification methods submitted for AOAC *Official Methods of Analysis*SM (OMA) status and/or for *Performance Tested Methods*SM (PTM) status. The requirements for single-laboratory validation (SLV) studies, independent validation studies, and collaborative validation studies for those methods are described.

2 Applicability

These guidelines are intended to be applicable to the validation of all candidate botanical identification methods (*Annex A*) submitted to AOAC for (*1*) OMA status through either a collaborative study or an alternative pathway study or (*2*) PTM certification.

3 Terms and Definitions

3.1 Botanical

Of, or relating to, plants or botany. May also include algae and fungi. May refer to the whole plant, a part of the plant (e.g., bark, woods, leaves, stems, roots, rhizomes, flowers, fruits, seeds, etc.), or an extract of the parts.

3.2 Botanical Identification Method (BIM)

A method that establishes identity specifications for a botanical material and determines, within a specified statistical limit, a binary test result: YES, the test material is a true example of the target botanical material and meets the identity specifications, or NO, it is not the target botanical. Thus, a BIM answers the question, "Is the test material the same as the target material?" not "What is this material?" In most cases, the method will achieve this goal by comparison of the test material with materials from the inclusivity panel and will return a YES/NO (or, in some cases, a consistent/nonconsistent) answer.

3.3 Candidate Method

The method to be validated or submitted for validation (*Annex A*).

3.4 Exclusivity

Ability of a BIM to correctly reject nontarget botanical materials.

3.5 Exclusivity Sampling Frame (ESF)

A list of practically obtainable nontarget botanical materials that have taxonomic, physical, or chemical composition characteristics similar to the target botanical and must give a negative result when tested by the BIM.

3.6 Exclusivity Panel

A subset of the ESF that is selected for the validation study. The identity of these materials should be verified by an appropriate method or process.

3.7 Identity Specification (IS)

The morphological, genetic, chemical, or other characteristics that define a target botanical material. Specifications may include, but are not limited to, data from macroscopic, microscopic, genetic (e.g., DNA sequencing), chromatographic fingerprinting (e.g., capillary electrophoresis, gas chromatography, liquid chromatography, or thin-layer chromatography), and spectral fingerprinting (e.g., infrared, near-infrared, nuclear magnetic resonance, ultraviolet/visible absorbance, or mass spectrometry) methods.

3.8 Inclusivity

Ability of a BIM to correctly identify variants of the target material that meet the identity specification.

3.9 Inclusivity Sampling Frame (ISF)

A list of practically obtainable botanical materials that are expected to give a positive result when tested by the BIM. The inclusivity frame should be sufficiently large that the botanical variation is adequately represented. Sources of variation may include, but are not limited to, species, subspecies, cultivar, growing location, growing conditions, growing season, and post-harvest processing.

3.10 Inclusivity Panel

A subset of the ISF that is selected for the validation study. These materials should be authenticated by an appropriate method.

3.11 Laboratory Sample

Sample as prepared for sending to the laboratory intended for inspection or testing.

3.12 Nontarget Botanical Material

Any botanical material that does not meet the identity specification.

3.13 Physical Form

Botanical materials exist in a number of physical forms. The form(s) will be specified by the standard method performance requirements (SMPRs).

3.14 Probability of Identification (POI)

The expected or observed fraction of test portions at a given concentration that give a positive result when tested by the BIM. A general description is provided in *Annex B*.

3.15 Sample

A small portion or quantity, taken from a population or lot that is ideally a representative selection of the whole. Sample homogeneity is usually determined with multiple samples.

3.16 Specified Inferior Test Material (SITM)

A botanical material mixture that has the maximum concentration of target material that is considered unacceptable, as specified by the SMPRs. The BIM must reject this material with a specified minimum level of (1 – POI) with 95% confidence. The ideal BIM would reject the SITM 100% of the time (i.e., accept 0% of the time). The SITM will typically be high-quality target material mixed with the worst-case (for identification) nontarget material.

3.17 Specified Superior Test Material (SSTM)

A botanical material mixture that has the minimum acceptable concentration of the target material, as specified by the SMPR. The BIM must identify this material with a specified minimum level of POI with 95% confidence. The ideal BIM would accept the SSTM 100% of the time. The SSTM will typically be high-quality target material mixed with a small amount of worst-case (for identification) nontarget material.

3.18 Standard Method Performance Requirements (SMPRs)

Performance requirements based on the fitness-for-purpose statement for each method. For BIMs, the SMPRs should include the physical form of the sample, the ISF, the ESF, the SSTM, the SITM, the number of samples for the inclusivity/exclusivity panels, and the desired probability and confidence limits for the method.

3.19 Target Botanical Material

The botanical material of interest as described in the identity specification.

3.20 Test Portion

The portion of the laboratory sample that is subjected to analysis by the method.

4 Validation Study Guidelines

A validated BIM requires a method validation study that demonstrates its acceptability according to the SMPRs. The guidelines presented here are intended to be applied to any qualitative BIM that returns a binary, YES/NO test result (*Annex A*). The guidelines provide technical guidance in validating the method based on the POI model (*Annex B*).

4.1 SMPRs

The SMPRs will be prepared by the appropriate AOAC body as per AOAC policy. The SMPRs will specify (*1*) the target botanical material, (*2*) the physical form of the material, (*3*) a list of botanical materials for the ISF/ESF, (*4*) composition of the SSTM and SITM, (*5*) maximum POI for the SITM and minimum POI for the SSTM, and (*6*) the desired probability and confidence limits for the inclusivity/exclusivity and SSTM/SITM measurements.

The SMPRs will consider the nature of the material being tested and determine the necessary breadth and depth of the inclusivity and exclusivity panels. In some cases, a few, very similar exclusivity panel materials may require in-depth testing (more test portions of a smaller group of materials). Conversely, the nature of the material may require greater breadth (fewer test portions of a greater number of materials).

The number of test portions needed should be determined on sound statistical grounds (*Annex C*) and subject matter expertise.

4.2 SLV Study

4.2.1 Scope

An SLV study is intended to determine the performance of a candidate method (*Annex A*). For validation purposes, the candidate BIM may be regarded as a black box providing a binary, YES/NO test result. The study is designed to evaluate performance parameters for the candidate method including (*1*) inclusivity/exclusivity, (*2*) POI for the SSTM and the SITM, and (*3*) POI as a function of the concentration of the target material (analytical response curve). This last parameter may be optional as specified by the SMPRs.

4.2.2 Inclusivity/Exclusivity Study

The purpose of this study is to confirm the ability of the candidate method to provide positive results (YES answers) for botanical materials on the inclusivity panel and negative results (NO answers) for materials on the exclusivity panel.

4.2.2.1 Inclusivity/Exclusivity Panel Selection

Botanical materials selected from the ISF/ESF will comprise the inclusivity/exclusivity panels. If the ISF/ESF specified by the SMPRs are sufficiently large, a representative subgroup will be selected for the panels by the method validator. Primary requirements for the panel materials are their availability and identity verification by an appropriate method or process. All test portions should be as uniform and homogeneous as possible. The level of replication of the inclusivity/exclusivity panels will be specified in the SMPRs.

4.2.2.2 Study Design

Prepare the test samples in a form appropriate for the candidate method. All test samples will be blinded and randomized so that the analyst(s) cannot know the identity of the samples. Analyze the test samples following the instructions of the candidate method.

4.2.2.3 Data Analysis and Reporting

The data will be analyzed for positive and negative responses. Unexpected results will be investigated, evaluated, and resolved prior to continuing the validation. The data is reported for individual inclusivity/exclusivity material as the number correctly identified. For example, "Of the 30 specific botanical materials of the inclusivity panel that were tested, 28 were identified correctly (gave a positive result) and two were not identified correctly (gave a negative result). Those materials not identified correctly were the following: …" or "Of the 30 specific botanical materials of the exclusivity panel that were tested, 27 were identified correctly (gave a negative result) and three were not identified correctly (gave a positive result). Those not identified correctly were the following: …" The study report should include a table titled "Inclusivity/Exclusivity Panel Results," which lists all materials tested, their source, origin, and essential characteristics and testing outcome. The implications of each unexpected result should be discussed and evaluated.

4.2.3 SSTM/SITM Study

The purpose of this study is to demonstrate method performance at two concentrations, the SSTM and the SITM.

4.2.3.1 Test Samples

The appropriate amount of a target material is selected from the inclusivity panel and is mixed with an appropriate amount of a nontarget material from the exclusivity panel to produce the SSTM and SITM as specified by the SMPRs. The test materials may be prepared using individual botanical materials from the inclusivity/exclusivity panels or composites of materials from the two panels as specified by the SMPRs.

All test portions should be as uniform and homogeneous as possible. The level of replication of the SSTM and SITM will be specified in the SMPR.

4.2.3.2 Study Design

Prepare the test samples in a form appropriate for the candidate method. All test samples will be blinded and randomized so that the analyst(s) cannot know the identity of the samples. Analyze the test samples following the instructions of the candidate method.

4.2.3.3 Data Analysis and Reporting

The data will be analyzed for positive and negative responses. For the SSTM and the SITM, report the POI results with 95% confidence intervals and the total number tested and the total number correctly identified. Comparison to SMPRs should be made and discussed.

4.2.4 Analytical Response Curve

This study will characterize the POI curve for mixtures of SSTM and SITM.

4.2.4.1 Test Samples

The appropriate amount of a target material is selected from the inclusivity panel and is mixed with an appropriate amount of a nontarget material from the exclusivity panel to produce mixtures with concentrations intermediate between the SSTM and SITM. The test materials shall be prepared using the same target and nontarget botanical material samples used in the SSTM and SITM study. The test materials may also be prepared by mixing appropriate ratios of the SSTM and SITM.

4.2.4.2 Study Design

Prepare the test samples in a form appropriate for the candidate method. All test samples will be blinded and randomized so that the analyst(s) cannot know the identity of the samples. Analyze the test samples following the instructions of the candidate method.

4.2.4.3 Data Analysis and Reporting

The data will be analyzed for positive and negative responses. For each mixture, report the POI results with 95% confidence intervals, the total number of samples tested, and the total number of positive responses. Plot the POI curve and confidence intervals.

4.3 Independent Validation Study

This study is identical to the SLV Study in Section 4.2.

4.4 Collaborative Study

The collaborative study is a route to an *Official Method*℠. The purpose of the collaborative study is to estimate the reproducibility and determine the performance of the candidate method among collaborators.

4.4.1 Number of Collaborators

A minimum of 10 independent laboratories reporting valid data is required. The study director should plan on including additional laboratories in the case of invalid data sets.

4.4.2 Number of Tests

Each collaborator receives 12 replicates of each material to be studied. At a minimum these materials will include the SSTM and SITM. Prepare the test samples in a form appropriate for the candidate method. All test samples will be blinded and randomized so that the analyst(s) cannot know the identity of the samples. Analyze the test samples following the instructions of the candidate method.

4.4.3 Data Analysis and Reporting

The data will be analyzed by the laboratory for positive and negative responses. For the SSTM and the SITM, report the POI results with confidence intervals for each laboratory, and for the

combined results. Estimate reproducibility as in *Annex C* and evaluate compared to the SMPRs.

ANNEX A
Candidate Method (or Prevalidation Study)

1 Scope

The candidate method must measure appropriate characteristics that are suitable to the question being asked and that will meet predetermined SMPRs. The method may be based on new principles or modifications of an existing method. The identity specifications will be based on morphological, genetic, and/or chemical characteristics, or any other defining feature of the botanical material. The candidate method may use visual inspection, DNA sequencing, instrumental analysis, or any other appropriate measurement. The measured characteristics will collectively provide a single analytical parameter that will be used to determine the final YES or NO result. The analytical parameter may be based on the degree of similarity or the degree of difference of the test sample and the reference material.

2 Inclusivity/Exclusivity Panel Selection

The method developer will select representative botanical materials from the ISF and ESF for use as target and nontarget botanical materials, respectively, in development of the method. These materials must be authenticated by an appropriate method.

3 Analytical Parameter

The method developer will prepare all the botanical samples in a form appropriate for the candidate method. The developer will analyze the target and nontarget botanical materials using the candidate method and develop an analytical parameter that is suitable for distinguishing between the two sets of materials.

4 Probability of Identification (POI)

Target materials will be mixed with systematically increasing amounts of nontarget materials to produce a series of target materials whose concentrations range from 100% to a concentration below the minimum acceptable concentration specified by the SMPRs. The developer will analyze the target and diluted target materials using the candidate method and determine the analytical parameter for each concentration.

5 Specific Superior/Inferior Test Materials

Based on the analytical parameters measured for the diluted target materials, a threshold value will be established that will permit positive identification of the minimum acceptable concentration of the target material with the specified confidence (e.g. 95%). The developer will use the threshold to determine a POI for each concentration (*Annex B*). The POIs measured for each concentration will be used to construct the POI curve.

6 Data Analysis and Reporting

The method developer will document the candidate method and the POI results.

ANNEX B
Understanding the POI Model

[*See Official Methods of Analysis* (2012) *Appendix K*, Part III, "Probability of Identification: A Statistical Model for the Validation of Qualitative Botanical Identification Methods," by Robert LaBudde and James M. Harnly, *J. AOAC Int.* **95**, 273–285 (2012). http://dx.doi.org/10.5740/jaoacint.11-266]

ANNEX C
Number of Test Portions

See Table C1.

Notes: (*1*) Enter the first column with the maximum error fraction tolerated by the SMPR, e.g., 10%.

(*2*) Select the sample size required by the number of misclassifications to be allowed, e.g., one erroneous result gives a sample size of $n = 48$ for a maximum error probability of 10%.

(*3*) Allowing more erroneous results increases the sample size required.

(*4*) The last (AOQL) column indicates the maximum error probability of a method which passes the SMPR for the test. For the example sampling plan indicated, this is 5.4%, approximately ½ of the maximum error probability in the SMPR. Typically the AOQL must be only 50–60% of the SMPR value to reliably pass the validation test. Method developers should take this into account.

Sample Size Required for Proportion

ASSUME: 1. Binary outcome (occur / not occur).
2. Constant probability rho of event occurring.
3. Independent trials (e.g., simple random sample).
4. Fixed number of trials N.

INFERENCE: 95% confidence interval lies entirely at or BELOW specified maximum rho.
DESIRED: Sample size N needed.
NOTES: 1. Based on modified Wilson score 1-sided confidence interval.
2. AOQL = Average Outgoing Quality Level

Maximum Probability rho	Sample Size N	**Maximum** Number Events x	**Minimum** Number Non-events y	1-sided **Upper** Confidence Limit on rho	Expected Lower Confidence Limit on rho	Expected Upper Confidence Limit on rho	Effective AOQL rho
50%	3	0	3	47.4%	0.0%	56.1%	28.1%
50%	10	2	8	45.9%	5.7%	51.0%	28.3%
50%	20	6	14	48.4%	14.5%	51.9%	33.2%
50%	40	14	26	48.0%	22.1%	50.5%	36.3%
50%	80	32	48	49.2%	30.0%	51.0%	40.5%
45%	2	0	2	57.5%	0.0%	65.8%	32.9%
45%	10	1	9	34.8%	0.0%	40.4%	20.2%
45%	20	5	15	43.2%	11.2%	46.9%	29.0%
45%	40	12	28	42.9%	18.1%	45.4%	31.8%
45%	80	28	52	44.1%	25.5%	45.9%	35.7%
40%	5	0	5	35.1%	0.0%	43.4%	21.7%
40%	10	1	9	34.8%	0.0%	40.4%	20.2%
40%	20	4	16	37.8%	8.1%	41.6%	24.8%
40%	40	10	30	37.6%	14.2%	40.2%	27.2%
40%	80	24	56	39.0%	21.1%	40.8%	30.9%
35%	6	0	6	31.1%	0.0%	39.0%	19.5%
35%	10	1	9	34.8%	0.0%	40.4%	20.2%
35%	20	3	17	32.2%	5.2%	36.0%	20.6%
35%	40	9	31	34.9%	12.3%	37.5%	24.9%
35%	80	21	59	35.0%	17.9%	36.8%	27.3%
30%	7	0	7	27.9%	0.0%	35.4%	17.7%
30%	10	0	10	21.3%	0.0%	27.8%	13.9%
30%	20	2	18	26.2%	2.8%	30.1%	16.4%
30%	40	7	33	29.3%	8.7%	31.9%	20.3%
30%	80	17	63	29.6%	13.7%	31.4%	22.6%
25%	9	0	9	23.1%	0.0%	29.9%	15.0%
25%	10	0	10	21.3%	0.0%	27.8%	13.9%
25%	20	1	19	19.6%	0.0%	23.6%	11.8%
25%	40	5	35	23.5%	5.5%	26.1%	15.8%
25%	80	13	67	24.1%	9.7%	25.8%	17.8%
20%	11	0	11	19.7%	0.0%	25.9%	12.9%
20%	20	1	19	19.6%	0.0%	23.6%	11.8%
20%	24	1	23	16.7%	0.0%	20.2%	10.1%
20%	36	3	33	19.1%	2.9%	21.8%	12.4%
20%	40	3	37	17.3%	2.6%	19.9%	11.2%
20%	48	5	43	19.9%	4.5%	22.2%	13.3%
20%	60	6	54	18.2%	4.7%	20.1%	12.4%
20%	72	8	64	18.7%	5.7%	20.4%	13.1%
20%	80	10	70	19.8%	6.9%	21.5%	14.2%
15%	20	0	20	11.9%	0.0%	16.1%	8.1%
15%	24	0	24	10.1%	0.0%	13.8%	6.9%
15%	36	1	35	11.5%	0.0%	14.2%	7.1%
15%	40	2	38	14.0%	1.4%	16.5%	8.9%
15%	48	3	45	14.6%	2.1%	16.8%	9.5%
15%	60	4	56	14.0%	2.6%	15.9%	9.3%
15%	72	5	67	13.6%	3.0%	15.2%	9.1%
15%	80	6	74	13.9%	3.5%	15.4%	9.4%
10%	40	0	40	6.3%	0.0%	8.8%	4.4%
10%	48	1	47	8.8%	0.0%	10.9%	5.4%
10%	60	2	58	9.6%	0.9%	11.4%	6.1%
10%	72	3	69	10.0%	1.4%	11.5%	6.5%
10%	80	3	77	9.0%	1.3%	10.5%	5.9%
5%	60	0	60	4.3%	0.0%	6.0%	3.0%
5%	72	0	72	3.6%	0.0%	5.1%	2.5%
5%	80	0	80	3.3%	0.0%	4.6%	2.3%
5%	90	1	89	4.8%	0.0%	6.0%	3.0%

Table C1

PART III
Probability of Identification: A Statistical Model for the Validation of Qualitative Botanical Identification Methods

A botanical is an herbal material that is frequently used as an ingredient in a dietary supplement regulated in the United States under the Federal Food, Drug, and Cosmetic Act of 1938, as amended by the Dietary Supplement Health and Education Act of 1994 (1). More recently, current Good Manufacturing Practices for foods and dietary supplements (2) issued by the U.S. Food and Drug Administration has tasked manufacturers with establishing specifications and developing a QA program for all botanical ingredients. As a consequence, both processors of botanicals and regulators are interested in the verification of the identity of botanical materials. Thus, the development of reliable methods for the identification of botanical materials and minimum acceptable levels of contamination are critical.

A botanical identification method (BIM) is any qualitative method that reliably identifies a botanical material and returns a binary result of either 1 = "identified" or 0 = "not identified." The actual method used can be presumed unknown and a "black box" with respect to the protocols involved in the validation studies. The BIM must be validated in terms of inclusivity, exclusivity, probability of identification, robustness, reproducibility, repeatability, and other criteria.

The heart of the BIM is the probability of identification (POI) model. The POI model has been developed as a means of characterizing and validating the performance of a qualitative method based on simple statistics and associated confidence intervals (3, 4). Figure 1 (modified from ref. 3) shows a plot where the concentration of the target material increases towards the right while the concentration of a nontarget material increases to the left. The parameter of interest is the POI (the vertical axis), which is defined as the probability, at a given percentage of target material, of getting a positive response by the detection method. The positive response of the BIM indicates that the test material matches the target botanical material. While the plot in Figure 1 is symmetrical, POI plots are usually asymmetrical. The POI model is based on the probability of detection model which was developed for binary qualitative methods (3, 4).

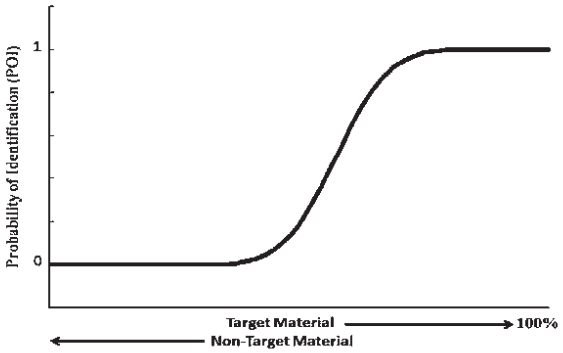


Figure 1. Probability of identification for botanical identification.

The POI, as illustrated in Figure 1, is dependent on the concentration of the target botanical material. The probability of a positive response increases as the concentration of the target botanical increases and decreases as the concentration of the nontarget material increases. The goal of method development and validation is primarily to determine if the method meets method performance requirements (MPRs), and secondarily to characterize how the method makes the transition from a negative to a positive response.

The MPRs, as established by the developer, will specify the target botanical materials (inclusivity sampling frame; ISF), the nontarget materials (exclusivity sampling frame; ESF), the physical form of the materials, the minimum concentration of target material that is acceptable in the presence of nontarget material, and the maximum concentration target material that is unacceptable. These latter materials are the specific superior and specific inferior test materials (SSTM and SITM, respectively). The idealized goal of the BIM is to discriminate (with a specified degree of confidence, e.g., 95%) between the SSTM (for which the POI is high) and the SITM (for which the POI is low). Additionally, samples of the SSTM and SITM may be mixed to obtain the intermediate test concentrations that are used to characterize the POI curve in its transitional range.

In some studies, full characterization of the transition of the POI curve may be of lesser importance and the intermediate concentrations omitted. In this care the only concentrations used are those for which the performance requirements are applied, typically the SITM and SSTM (0% and 100% SSTM, respectively). Two factors are important to method development: industrial-regulatory requirements, and the technological limit (state of the measurement art). If the technological limit exceeds the industry-regulatory requirement, then the industrial-regulatory requirement can be set at a value reasonably attainable by existing technology. In this case, the cost of the analysis may be the major factor governing validation study design. If the technological limit cannot meet the industrial-regulatory requirement, then improved technology must be developed before a BIM fit for the purpose intended can be found.

Glossary

Analytical parameter (AP).—A measured or computed analytical value used to determine whether the test material matches the target material. The analytical parameter may be based on morphological

---

A qualitative botanical identification method (BIM) is an analytical procedure that returns a binary result (1 = identified, 0 = not identified). A BIM may be used by a buyer, manufacturer, or regulator to determine whether a botanical material being tested is the same as the target (desired) material, or whether it contains excessive nontarget (undesirable) material. The report describes the development and validation of studies for a BIM based on the proportion of replicates identified, or probability of identification (POI), as the basic observed statistic. The statistical procedures proposed for data analysis follow closely those of the probability of detection (POD), and harmonize the statistical concepts and parameters between quantitative and qualitative method validation. Use of POI statistics also harmonizes statistical concepts for botanical, microbiological, toxin, and other analyte identification methods that produce binary results. The POI statistical model provides a tool for graphical representation of response curves for qualitative methods, reporting of descriptive statistics, and application of performance requirements. Single collaborator and multicollaborative study examples are given.

Reference: LaBudde, R.A., & Harnly, J.M. (2012) *J. AOAC Int.* **95**, 273–285. http://dx.doi.org/10.5740/jaoacint.11-266

The POI statistical model was approved by the AOAC Official Methods Board on October 13, 2011.

features, genetic sequences, chromatographic patterns, spectral patterns, or any other metric appropriate for the target material.

Botanical.—Of or relating to plants or botany. May also include algae and fungi. May refer to the whole plant, a part of the plant (e.g., bark, woods, leaves, stems, roots, rhizomes, flowers, fruits, seeds, extracts, etc.), or an extract of the plant.

BIM.—A method that establishes identity specifications for a botanical material and determines, within a specified statistical limit, a binary result: yes, the test material is a true example of the target botanical material and meets the identity specifications; or no, it is not the target botanical. Thus, a BIM answers the question, "*Is the test material the same as the target material?*" not "*What is this material?*" In most cases, the method will achieve this goal by comparison of the test material with materials from the inclusivity panel and will return a yes/no (or, in some cases, a consistent/nonconsistent) answer.

Candidate method.—The method to be validated.

Exclusivity.—Ability of a BIM to correctly reject nontarget botanical materials.

ESF.—A list of practically obtainable nontarget botanical materials that have similar taxonomic, physical, or chemical composition characteristics that are expected to give a negative result when tested by the BIM.

Exclusivity panel.—A subset of the ESF that is selected for the validation study. These materials should be authenticated by an appropriate method.

False-negative fraction (FNF).—1–POI for 100% SSTM. Not defined for other concentrations.

False-positive fraction (FPF).—POI for 100% SITM. Not defined for other concentrations.

Identity specification.—The morphological, genetic, chemical, or other characteristics that define a target botanical material. Specifications may include, but are not limited to, data from macroscopic, microscopic, genetic (e.g., DNA sequencing, barcoding), chromatographic fingerprinting (e.g., CE, GC, LC, TLC), and spectral fingerprinting (e.g., IR, NIR, NMR, MS, UV-Vis) methods.

Inclusivity.—Ability of a BIM to correctly identify variants of the target material that meet the identity specification.

ISF.—A list of practically obtainable botanical materials that are expected to give a positive result when tested by the BIM. The inclusivity sampling frame should be sufficiently large that the botanical variation is adequately represented. Sources of variation may include, but are not limited to, species, subspecies, cultivar, growing location, growing conditions, growing season, and post-harvest processing.

Inclusivity panel.—A subset of the ISF that is selected for the validation study. These materials should be authenticated by an appropriate method.

Laboratory sample.—Sample as prepared for sending to the laboratory intended for inspection or testing.

MPRs.—Performance requirements based on the fitness-for-purpose statement for each method. For BIMs, the MPRs should minimally include the physical form of the sample, the ISF, the ESF, the SSTM, and the SITM.

Nontarget botanical material.—Any botanical material that does not meet the identity specification.

Physical form.—Botanical materials exist in a number of physical forms. The form(s) to be analyzed by the method will be specified by the MPRs.

POI.—The expected or the observed fraction of test portions that provide a positive result at a given concentration when tested by the BIM.

Sample.—A small quantity, taken from a population or lot that is a representative selection of the whole.

SITM.—A mixture of botanical materials that contains the maximum concentration of target material that is considered unacceptable, as specified by the MPRs. The BIM must reject this material with a specified minimum level of (1–POI) with 95% confidence. The ideal BIM would reject the SITM 100% of the time (i.e., identify 0% of the time). The SITM will typically be high-quality target material mixed with worst-case (for identification) nontarget material.

SSTM.—A mixture of botanical material that contains the minimum acceptable concentration of the target material, as specified by the MPR. The BIM must identify this material with a specified minimum level of POI with 95% confidence. The ideal BIM would identify the SSTM 100% of the time. The SSTM will typically be high-quality target material mixed with a small amount of worst-case (for identification) nontarget material.

Target botanical material.—The botanical material of interest as described in the identity specification.

Target material concentration.—The percentage, by weight, of the target botanical material in the sample.

Test portion.—The portion of the laboratory sample that is subjected to analysis by the method.

Inclusivity Panel

When a botanical material is identified for development of a BIM, a target material is usually specified. Biological materials, however, are complex. While the genotype of a species or subspecies may be relatively stable, the phenotype (metabolite composition) will vary with location, season, weather, and many other variables. Thus, "target material" becomes "target materials." Ideally, the target materials will encompass the expected botanical variation.

An inclusive list of all the variations for a target material can be quite extensive and impractical. For example, the list for a specific botanical might ideally include samples from the last 10 years from eight international locations (80 samples). In reality, only 25 of the desired samples may be practically obtainable. These 25 obtainable samples comprise the ISF. Of these 25 samples, only 10 may be selected for method development/validation. These 10 samples comprise the inclusivity panel.

For each candidate BIM, the MPRs must provide a list of all necessary botanical variants that should provide a positive identification. This should include species, varieties, geographic or seasonal variants, and other variants that are believed to possibly associate with BIM identification performance. The information tabulated should include variety, season, locality, source from which the variant is obtainable, species, variety or subclass, and whether or not it is essential that the variant be tested. The age of the plant may also be a factor of importance. The subset of this list, which is practically obtainable for a validation study, is the ISF.

The MPRs should identify the minimum number of materials in the ISF that must be tested to verify identifiability (inclusivity panel), as well as the number of replicates needed. If at all possible, any exchangeability (choice among variants which MPRs do not discriminate) should result in random selection from the ISF.

Generally, the inclusivity panel of target variants should include all of the ISF if the number of variants is small. Otherwise, all

necessary variants plus additional ones randomly selected should comprise the inclusivity panel. More randomized replicate variants may allow a quantitative statistical inference to be made concerning inclusivity. An inclusivity panel with no randomization, only subjective selection, does not permit statistical statements of inference with respect to inclusivity.

Exclusivity Panel

The list of nontarget materials can be quite extensive, theoretically including all the botanicals not on the inclusivity list. However, of prime interest are those materials that might accidentally or intentionally be used to replace or augment the target materials. The exclusivity list should include botanical materials that are closely related taxonomically, morphologically, or phenotypically. Again, this list may be extensive and impractical. The ESF will comprise those botanical materials that are practically obtainable. The exclusivity panel will comprise those samples used for method development and validation.

The MPRs must provide a list of all necessary or commonly encountered nontarget botanical materials and variants. This list should include botanical materials that are believed to accidentally or intentionally alter the composition of the target material. The information tabulated should include variety, season, locality, source from which the variant is obtainable, species, variety or subclass, and whether or not it is essential that the nontarget material be tested. The subset of this list, which is practically obtainable for a validation study, should then be identified as the ESF.

The MPRs should identify the minimum number of nontarget materials of the ESF that should be included on the exclusivity panel and be tested to verify non-identifiability, as well as the number of replicates needed. If at all possible, any exchangeability (choice among variants which expertise does not discriminate) should result in random selection from the ESF.

Generally, the exclusivity panel of authentic variants should include all of the ESF if the number of variants is small. Otherwise, all necessary variants, plus optional ones randomly selected, should comprise a set as specified by the ERP. More replicates and randomization may allow a quantitative statistical inference to be made concerning exclusivity.

Inclusivity and Exclusivity Testing

The purpose of inclusivity/exclusivity testing is to verify that the BIM correctly identifies all of the botanical materials listed in the ISF and correctly rejects all nontarget materials listed in the ESF. The BIM should clearly and unequivocally discriminate between the target and nontarget materials. Testing materials from the inclusivity/exclusivity panels should provide sufficient confidence that this is the case. The number of samples tested and the number of replicates is specified by the MPRs.

Typically, inclusivity/exclusivity panel results are verified during method development. Any unexpected results should be followed up with a minimum number of additional replications (determined by the MPRs) to characterize the POI on the variant quantitatively. If the variant fails to meet minimum acceptable performance requirements as set by the MPRs, the exception should be noted in the study report and reviewed for acceptability by the relevant method reviewers.

If the method development results are acceptable, inclusivity and exclusivity should be verified in an independent laboratory, although possibly on a less-intensive (fewer replicates or randomly selected variants) basis, as the objective is verification, not validation. If no randomization is used, all that can be reported are the actual results obtained, but without suggestive quantitative statistics. For example, without randomization, the use of percentages or other quantitative measures is inappropriate.

Performance Requirements and the Specification and Preparation of the SITM and SSTM

After inclusivity and exclusivity studies have been completed, target and nontarget material(s) are chosen to verify that the method can discriminate between the SSTM and the SITM. Either the worst-case nontarget materials, or perhaps the most common nontarget materials, would typically be chosen. In addition, a combination of target and nontarget materials should be selected to challenge method performance (worst-case, most common, etc.). The number of samples tested and the number of replicates is specified by the MPRs.

The MPRs should identify the composition and the minimum POI acceptable (with 95% confidence) for the SSTM and SITM. The SSTM and SITM would be made of the target material(s) mixed with the combination of nontarget material(s).

Application of the POI to an Analytical Method

Analytically, a BIM will be based on a series of measured values. These values may be derived from morphological features, genetic sequences, chromatographic patterns, spectral patterns, or any other metric appropriate for the target material. These values will be combined to provide a single AP that will be used to determine whether the test sample does or does not match the materials from the inclusivity panel. This decision is made by comparing the AP of the test material to a threshold value that provides the level of identification specified by the MPRs.

The first step in the development of the method is the selection of the analytical approach and the analysis of samples from the ISF and ESF. Multiple replicates of multiple samples should, ideally, give results similar to those in Figure 2. Here, the AP, not the POI, is plotted on the vertical axis. The standard deviations (SDs) are shown as sample distribution functions, rather than as error bars. Ideally, the separation of the ISF and ESF samples should be as large as possible. For the data in Figure 2, the threshold to distinguish between the ISF and ESF can be placed at almost any value of the AP.

The width of the sample distribution function will depend on the number of samples analyzed from the ISF and ESF. If replicates

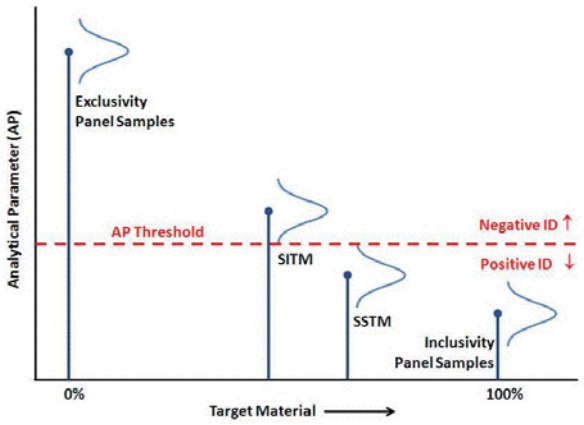


Figure 2. Inclusivity/exclusivity and SSTM/SITM characterization.

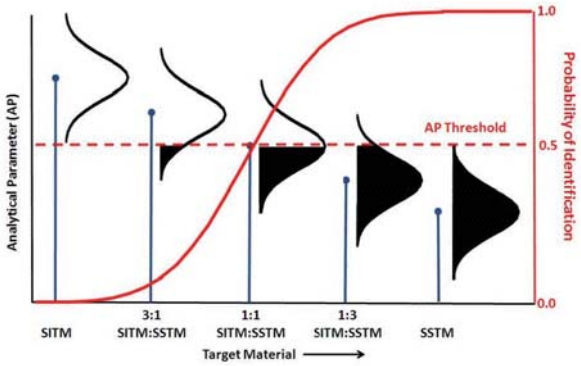


Figure 3. Conversion of SSTM, SITM, and intermediate concentrations to POI.

of a single sample are analyzed, then the width of the distribution will be narrow (a smaller SD), and only reflect the instrumental variance. As more samples are analyzed from the ISF and ESF, the distribution functions will broaden, reflecting the increasing biological variance.

The next step is to determine whether the method can distinguish between the SSTM and the SITM. The concentrations of the SSTM and the SITM are specified by the MPRs. Figure 2 illustrates an arbitrary specification. It can be seen that the distributions of the SSTM and SITM are completely resolved and the threshold must be located exactly between the two distributions to provide 100% identification of the SSTM (POI = 1) and 100% rejection of the SITM (POI = 0). If the concentration of target material in the SSTM was lower, or the concentration in the SITM higher, the distribution functions would overlap and 100% identification or rejection would not be possible. In this case, the confidence limit would have to be lowered or another method selected.

Finally, the shape of the POI curve can be determined. As shown in Figure 3, concentrations of the target materials that fall between the SSTM and SITM must be prepared. In each case, the threshold will intersect each peak and determine the POI. As the SSTM:SITM values change from 1:0 to 3:1 to 1:1 to 1:3 to 0:1, the POI decreases from 1.0 to 0.9 to 0.5 to 0.1 to 0.0.

The models in Figures 2 and 3 assume that the SITM and SSTM have the same, symmetrical distribution function and width. This is not a reasonable assumption for real samples. However, the POI model is valid regardless of the shape of the distribution functions involved.

A Specific Example: American Ginseng Mixed with Asian Ginseng

The data set presented here illustrates the analytical measurements discussed in the previous section. The target botanical material is American ginseng (AG) and the nontarget material is Asian ginseng (CG). The inclusivity panel consists of 43 AG samples grown in the United States (harvested over 3 years from 20 different farms in Wisconsin), and the exclusivity panel consists of eight CG samples grown in China (Table 1).

The AG and CG samples were analyzed by direct injection MS, and yielded spectra with approximately 1000 ions. The SSTM and SITM were generated synthetically by combining different percentages of the AG and CG mass spectra. For example, the spectra for 98% AG mixed with 2% CG was computed as 0.98 of an AG spectra added to 0.02 of a CG spectra. In all, 344 SSTM spectra were generated (43 AG × 8 CG).

The multivariate data set (395 samples × 1000 variables) was analyzed using soft independent modeling of class analogy (SIMCA; *Annex A*). SIMCA fit a principal component model to the data for the inclusivity panel (100% AG) and produced a goodness-of-fit value, the Q residual, for every sample analyzed. The Q residual was used to compare the test (100% CG, SSTM, and SITM) and the target (100% AG) materials. In every case, the SIMCA model was based on 100% AG and a single principal component. The Q residual describes how far a sample falls outside the model (*Annex A*).

Figure 4 (A) shows the inclusivity/exclusivity study. The Q residual is plotted for individual samples. With 100% AG

Table 1. Panax samples analyzed in this study

No.	Label	Provider	Source
	Inclusivity panel (American ginseng)		
26	American ginseng		USA
13	American ginseng		USA
4	American ginseng		USA
	Exclusivity panel (Chinese ginseng)		
3	Asian ginseng, red	American Herbal Pharmacopoeia 2	China
1	Kirin Red No. 1	Internet retailer	China
1	Kirin Red No. 3	Internet retailer	China
1	Kirin Red No. 5	Internet retailer	China
1	Shih Chu No. 25	Internet retailer	China
1	Shih Chu No. 80	Internet retailer	China
	SSTM/SITM[a]		
344	SSTM[a]	0.98 American ginseng + 0.02 Asian ginseng	
344	SITM[a]	0.90 American ginseng + 0.10 Asian ginseng	

[a] In each case, each of the 43 American ginseng samples were mixed with each of the eight Asian ginseng samples (43 × 8 = 344).

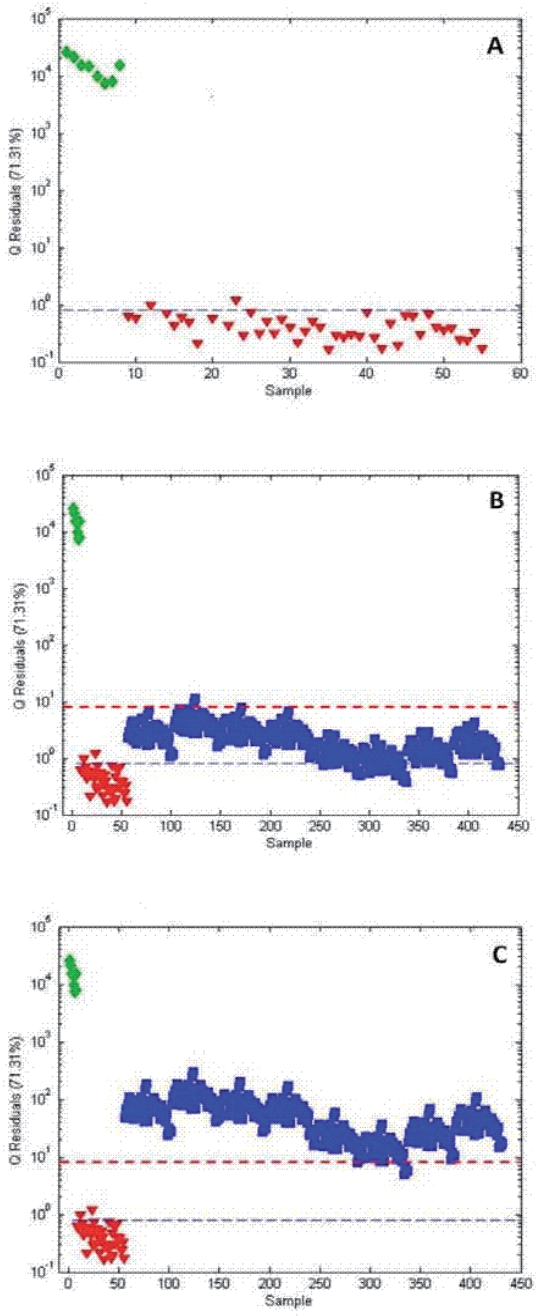


Figure 4. SIMCA plots for (A) 100% American ginseng (AG; ▼) and 100% Asian ginseng (CG; ◆); (B) SSTM (■), 100% AG, and 100% CG; and (C) SITM (■), 100% AG, and 100% CC.

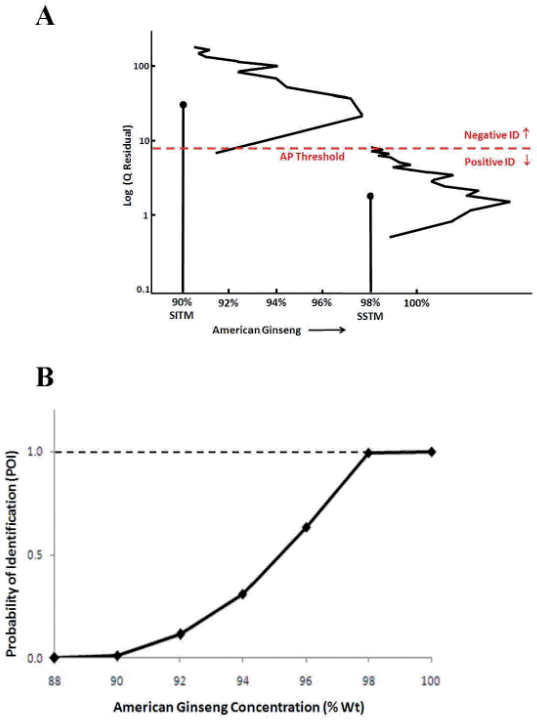


Figure 5. Target material AG, nontarget material CG: (A) SITM and SSTM, and (B) POI.

(inclusivity panel samples) as the model, the CG (exclusivity panel samples) falls well above the 95% confidence limit (dashed line). Both the AG and CG show considerable variation on the vertical axis, which reflects biological variation. Two of the AG samples fall above the 95% confidence limit, which is 4.6% for 43 samples and is to be expected.

For the SSTM/SITM study, 98 and 90% AG were arbitrarily selected as the MPRs for this model. Figure 4 (B) shows the SSTM samples (98% AG), as well as 100% AG and 100% CG samples. The pattern of eight groupings for the SSTM samples reflects that all 43 AG samples were diluted by each of the eight CG samples in sequence. A threshold of a Q residual value of 9.0 was selected arbitrarily and provides 99.4% positive identification (342 out of 344).

Figure 4 (C) shows the SITM at 90% AG. The threshold provides negative identification of the SITM for 99.1% of the samples (341 out of 344). The distribution of the SSTM and SITM are plotted in Figure 5 (A). The distributions appear to be roughly symmetrical. However, since the vertical axis is a logarithmic scale, the distributions are badly skewed on a linear scale and have dramatically different widths. If the SSTM were specified at a lower concentration of AG, or the SITM at a higher concentration, the method would not be appropriate unless lower confidence limits were chosen.

Based on the AP threshold shown in Figures 4 (B, C) and 5, the POI in Figure 5 (B) was computed. Synthetic samples of 96, 94, and 92% were generated and analyzed. The curve shape for the POI is very non-symmetric.

For our example, the SSTM corresponds to 98% AG mixed with 2% CG. The required minimum POI is 0.90, with 95% confidence for 100% SSTM (Table 2). The SITM corresponds to 90% AG mixed with 10% CG. The required maximum POI is 0.10,

Table 2. Example performance requirements

Requirement	SSTM, %	Measure	Limit	No. of replicates to be tested	No. of failures allowed[a]
POI	100	95% 1-sided LCL	0.90 (FNF<0.10)	60	2
POI	0	95% 1-sided UCL	0.10 (FPF<0.10)	60	2

[a] In each case, no more than two failures are allowed.

with 95% confidence. Table 2 shows that, for these performance requirements, 60 replicates must be tested at each level with no more than two failures. More stringent requirements (i.e., 0.95 and 0.05, with 95% confidence) would require more replicates and/or fewer failures. Conversely, less-stringent requirements would require fewer replicates. Depending upon the desired performance requirement for SSTM or SITM, alternative test plans (confidence levels) may be selected from Table 3. For more plans, *see* LaBudde (5).

Single-Laboratory Validation

Consider an example of a BIM being evaluated with respect to the performance requirements of Table 2. The internal operating methodology of the BIM is possibly a trade-secret of the method developer, and may not be known at the time of validation. All that is known for sure is that a test portion is utilized by the method, and binary result of yes = Identified or no = Not Identified is returned.

Consider testing in a single independent laboratory, or an SLV. With respect to the performance requirements of Table 2, the SITM and SSTM are used to prepare mixtures in the proportions 0:100%, 33:67%, 67:33%, and 100:0%. From each of these mixtures, 60

Table 3. Alternative test plans to obtain 1-sided upper 95% modified Wilson confidence limit at or below specified maximum value for FNF or FPF[a]

Specified maximum[b]	No. of replicates to be tested	No. of failures allowed[c]	1-sided 95% UCL[d]	2-sided 95% LCL[e]	2-sided 95% UCL[e]	AOQL[f]
0.20	11	0	0.197	0.000	0.259	0.129
0.20	20	1	0.196	0.000	0.236	0.118
0.20	24	1	0.167	0.000	0.202	0.101
0.20	36	3	0.191	0.029	0.218	0.124
0.20	48	5	0.199	0.045	0.222	0.133
0.20	72	8	0.187	0.057	0.204	0.131
0.15	20	0	0.119	0.000	0.161	0.081
0.15	24	0	0.101	0.000	0.138	0.069
0.15	36	1	0.115	0.000	0.142	0.071
0.15	48	3	0.146	0.021	0.168	0.095
0.15	72	5	0.136	0.030	0.152	0.091
0.10	40	0	0.063	0.000	0.088	0.044
0.10	48	1	0.088	0.000	0.109	0.054
0.10	60	2	0.096	0.009	0.114	0.061
0.10	72	3	0.100	0.014	0.115	0.065
0.05	60	0	0.043	0.000	0.060	0.030
0.05	72	0	0.036	0.000	0.051	0.025
0.05	96	1	0.045	0.000	0.057	0.028
0.02	130	0	0.020	0.000	0.029	0.014
0.02	240	1	0.018	0.000	0.023	0.012
0.01	280	0	0.010	0.000	0.014	0.007

[a] Excerpted from LaBudde (5).

[b] Desired maximum level of FNF or FPF to attain with 95% confidence.

[c] Maximum number of failures that can occur in the replicates tested and still meet specification.

[d] Worst-case 1-sided 95% modified Wilson upper confidence limit on FNF or FPF if maximum failures are observed.

[e] 95% modified Wilson 2-sided confidence interval on FNF or FPF if maximum failures are observed.

[f] Observed FNF or FPF corresponding to maximum failures allowed.

Table 4. Observed SLV results for example BIM

SSTM, %	No. of test portions	No. identified	No. not identified	POI
0.0	60	1	59	0.0167
33.3	60	7	53	0.1167
66.7	60	27	33	0.4500
100.0	60	60	0	1.0000

test portions are prepared, randomized, and labeled in a masked way. The test portions are measured by the BIM, each with a result of 0 or 1. Suppose example results are as shown in Table 4. Note the FPF performance requirement succeeds at 0% SSTM, because no more than two test portions reported identification. Also, the FNF performance requirement at 100% SSTM succeeds because, in both cases, fewer than two test portions were not identified.

Using the methods of Wehling et al. (3) and LaBudde (6, 7), the reported 1-sided and 2-sided 95% confidence intervals on the POI would be as shown in Table 5. Note that the 1-sided 95% confidence limit for the POI falls below 10% at 0% SSTM, and above 90% at 100% SSTM, indicating performance requirement success. The results in Table 5 are plotted in Figure 6.

Because the concentrations (% SSTM) are known with certainty here, one of several regression models might be fit to possibly obtain more precise estimates of POI and its confidence limits (although this is not guaranteed), but at the expense of some additional assumptions (*see Annex B*).

Collaborative Study

The primary purpose of a collaborative study is to establish that performance is reproducible among different collaborators (laboratories). A secondary purpose might be to compare the candidate method to another (possibly gold standard) method to establish differential performance (e.g., equivalency) across laboratories.

The primary purpose requires a minimum number of collaborators whose data persist (i.e., not excluded for cause) until the final results of the study. Rules of thumb in statistical mixed modeling (treating the collaborator effect as random) suggest that fewer than six collaborators does not allow inference with respect to the general collaborator population, eight collaborators allows reasonable estimation, and 10 collaborators is desirable. More than 10 collaborators is useful, but not necessary. For fewer than six collaborators, the collaborator effect should be regarded as fixed, and any inferences are applicable only to that particular set of collaborators, not some hypothetical general population of collaborators. The recommendation, therefore, is that 12 or more collaborators should be enrolled in the study, with a desired 8 to 10 remaining after removal for cause, and an absolute limit of no fewer than six remaining until the study end. Studies with this minimum number of collaborators can hope to provide a measure of collaborator effect or collaborator-method interaction, if one of reasonably large size exists.

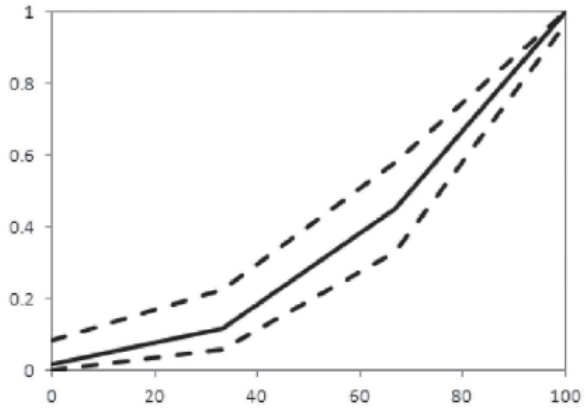


Figure 6. Expected POI versus %SSTM for an example BIM showing POI (solid line), lower 95% confidence limit (dashed line below the POI), and upper 95% confidence limit (dashed line above the POI). Note the POI at 0% is the false-positive fraction and 1-POI at 100% is the false-negative fraction.

Concentration levels (i.e., percentage of SSTM in a SSTM:SITM mixture) must include 0% SSTM (100% SITM) and 100% SSTM (0% SITM) in order to establish performance requirements (Figure 2). In addition, it is sometimes beneficial to provide for two intermediate concentrations (e.g., 33 and 67%) in order to provide information about identification performance across the range where the POI changes.

In order to isolate a collaborator effect in the presence of quantal noise (repeatability error), 12 replicates per collaborator is the suggested minimum. Therefore, the smallest acceptable collaborative study final data would be six collaborators × 12 replicates = 72 test portions.

It should be noted that due to the intercollaborator variation, a performance requirement imposed on a collaborative study will be more difficult for a candidate BIM to achieve than that imposed on an SLV study with the same number of total replicates. The performance requirements imposed on a single laboratory study and a collaborative study should be logically and statistically consistent.

The study director could, for example, prepare batches of SITM and SSTM, then prepare samples of mixtures at the 0:100%, 33:67%, 67:33%, and 100:0% proportions. From each of the well-mixed sample aliquots, test portions would be selected, such that each participating collaborator would receive the requisite number

Table 5. Reported SLV results

SSTM, %	*n*	ID	Not ID	POI	1-sided 95%	LCL 95%	UCL 95%
0.0	60	1	59	0.0167	0.0713	0.0000	0.0886
33.3	60	7	53	0.1167		0.0577	0.2218
66.7	60	27	33	0.4500		0.3309	0.5751
100.0	60	60	0	1.0000	0.9568	0.9398	1.0000

of replicates (*see* section on SLV). All test portions for each collaborator would be randomly assigned IDs before distribution. The study is masked so that collaborators cannot visually identify the composition of the test portions. Additional unmasked test portions may be provided for proficiency training purposes. Each collaborator would use the BIM according to instructions to analyze each test portion provided, and report results by test portion number and 1 = Identified or 0 = Not Identified.

Suppose a collaborative study is to be evaluated with respect to the performance requirements of Table 2. The primary goal is to validate that performance is sufficiently homogeneous across collaborators and that the performance requirements are met. As mentioned before, the number of replicate test portions for each collaborator should be 12 or more to control the quantal repeatability error sufficiently to allow detection of an intercollaborator effect. Suppose the plan was to enroll 12 collaborators, with the expectation that on or two might have to be removed for cause (spoilage of test portions, failing to follow instructions, cross-contamination, etc.) Consequently 144 test portions are prepared for each of the four % SSTM values (0, 33.3, 66.7, and 100%).

After completion of the study, two collaborators are removed for cause, and the results shown in Table 6 are obtained. For the 0% SSTM concentration, the statistical analysis of the data gives the results in Table 7. There is no detected intercollaborator effect (P-value = 0.43, point estimate = 0.00, confidence interval includes 0.000 and has an upper limit of 0.040), and the upper 2-sided confidence limit for combined POI is 0.0457, well below the performance requirement of 0.10. There is little evidence that the method is irreproducible, and the method meets the POI (or FPF) performance requirement.

For the 33% SSTM concentration, the statistical analysis of the data gives the results in Table 8. Again, there is no detected intercollaborator effect (P-value = 0.66), so there is little evidence that the method is irreproducible.

For the 67% SSTM concentration, the statistical analysis of the data gives the results in Table 9. Once again, there is no detected intercollaborator effect (P-value = 0.18), so there is little evidence that the method is irreproducible.

Finally, for the 100% SSTM concentration, the statistical analysis of the data gives the results in Table 10. There is no detected intercollaborator effect (P-value = 0.25, point estimate = 0.027, confidence interval includes 0.000 and has an upper limit of 0.093), and the lower 2-sided confidence limit for combined POI is 0.917, well above the performance requirement of 0.90. There is little evidence that the method is irreproducible, and the method meets the POI (or FNF) performance requirement.

Lot-Lot Variability, Time Stability, and Robustness Studies

The SLV and collaborative studies discussed above do not represent worst-case, end-of-life conditions with respect to method materials and parameters. For this reason, it is customary to augment these studies with additional studies to verify proper results despite reasonable variations among method materials, equipment, and parameters.

A lot-lot variability study is meant to verify results across different lots of method materials (supplies used) and sets of equipment. Each lot would consist of a different manufactured or prepared batch of materials (reagents, supplies, etc.), and possibly a different set of measurement equipment. Date of manufacture is not an issue in this study, only variation among lots, so ideally, the lots tested should have been produced at near the same times.

Table 6. Collaborative study results

SSTM, %	Collaborator	Replicates	No. identified
0	1	12	1
0	2	12	0
0	3	12	0
0	4	12	0
0	5	12	0
0	6	12	0
0	7	12	0
0	8	12	0
0	9	12	0
0	10	12	0
33.33	1	12	2
33.33	2	12	2
33.33	3	12	2
33.33	4	12	2
33.33	5	12	0
33.33	6	12	1
33.33	7	12	1
33.33	8	12	4
33.33	9	12	2
33.33	10	12	3
66.67	1	12	4
66.67	2	12	9
66.67	3	12	5
66.67	4	12	8
66.67	5	12	7
66.67	6	12	4
66.67	7	12	7
66.67	8	12	3
66.67	9	12	8
66.67	10	12	5
100	1	12	12
100	2	12	10
100	3	12	11
100	4	12	12
100	5	12	12
100	6	12	11
100	7	12	12
100	8	12	12
100	9	12	12
100	10	12	12

Table 7. Collaborative study results for 0% SSTM concentration

AOAC Binary Data Interlaboratory Study Workbook Study Reported Values, Version 2.2					
Sample ID 0% SSTM					
Sequence	Item	Symbol	Value	Approximately 95% LCL[a]	Approximately 95% UCL[b]
1	Total number of laboratories	p	10		
2	Total number of replicates	Sum(n(L))	120		
3	Overall mean of all data (grand mean)	LPOI or LPOD	0.0083	0.0015	0.0457
4	Repeatability SD	s(r)	0.0913	0.0807	0.1713
5	Among-laboratories SD	s(L)	0.0000	0.0000	0.0402
6	Homogeneity test of laboratory PODs	P-value	0.4303		
7	Reproducibility SD	s(R)	0.0913	0.0814	0.1064
8	Intraclass correlation coefficient for repeatability	I(r)	1.0000	0.8335	1.0000

[a] LCL = Lower confidence level.

[b] UCL = Upper confidence level.

Just as with collaborators in a collaborative study, estimation of the lot random effect requires that at least six different lots be involved in the study. Each lot should result in attainment of any BIM performance requirements, and the variation in performance among lots should be immaterial in size.

A time stability study is meant to verify that there is no material degradation in performance over the life of lots of materials and equipment. This may be accomplished by determination of the parametric aging effect by use of time-staggered lots, or simply verifying performance on end-of-life lots.

Note that the lot-lot variability and time-stability studies cannot be merged into a single study unless there are sufficient replicate lots at or near the same time point(s) to allow separation of the lot-lot and time effects. If lot-lot and time effects are negatively correlated, one factor may mask the effect of the other in an inadequate combined study (e.g., a different single lot at each different time point). Testing only end-of-life lots would be a satisfactory combined study, even though time and lot effects could not be resolved.

A robustness study (also denoted a sensitivity study) is meant to verify performance under worst-case conditions of method critical parameter (e.g., times, temperatures, concentrations) variation. Disturbances of method parameters should reflect maximum excursions to be expected in practical use. Performance requirements should be met at each of these excursions. The statistical design should be capable of measuring at least main effects.

Conclusions

The purpose of a qualitative BIM is to discriminate between acceptable target material and target material with an unacceptable concentration of nontarget material. This concept was particularized to discrimination between the SSTM and SITM for the purpose of method validation. A general overview of the application of the POI model and analysis was given, which allows validation and/or characterization of qualitative BIMs. Examples are given for both SLV and collaborative studies with MPRs. The use of POI statistics harmonizes statistical concepts among botanical, microbiological, toxin, and other analyte identification or detection methods for which binary results are obtained. The POI statistical model provides a tool for graphical representation of response curves for qualitative methods, reporting of descriptive statistics, and application of performance requirements.

Table 8. Collaborative study results for 33.33% SSTM concentration

AOAC Binary Data Interlaboratory Study Workbook Study Reported Values, Version 2.2					
Sample ID 33.33% SSTM					
Sequence	Item	Symbol	Value	Approximately 95% LCL	Approximately 95% UCL
1	Total number of laboratories	p	10		
2	Total number of replicates	Sum(n(L))	120		
3	Overall mean of all data (grand mean)	LPOI or LPOD	0.1583	0.0913	0.2253
4	Repeatability SD	s(r)	0.3703	0.3272	0.4266
5	Among-laboratories SD	s(L)	0.0000	0.0000	0.1400
6	Homogeneity test of laboratory PODs	P-value	0.6563		
7	Reproducibility SD	s(R)	0.3703	0.3304	0.4275
8	Intraclass correlation coefficient for repeatability	I(r)	1.0000	0.8889	1.0000

Table 9. Collaborative study results for 66.67% SSTM concentration

AOAC Binary Data Interlaboratory Study Workbook Study Reported Values, Version 2.2					
Sample ID 66.67% SSTM					
Sequence	Item	Symbol	Value	Approximately 95% LCL	Approximately 95% UCL
1	Total number of laboratories	p	10		
2	Total number of replicates	Sum(n(L))	120		
3	Overall mean of all data (grand mean)	LPOI or LPOD	0.5000	0.3919	0.6081
4	Repeatability SD	s(r)	0.4939	0.4364	0.5222
5	Among-laboratories SD	s(L)	0.0948	0.0000	0.2779
6	Homogeneity test of laboratory PODs	P-value	0.1783		
7	Reproducibility SD	s(R)	0.5029	0.4489	0.5222
8	Intraclass correlation coefficient for repeatability	I(r)	0.9644	0.7547	1.0000

Table 10. Collaborative study results for 100.0% SSTM concentration

AOAC Binary Data Interlaboratory Study Workbook Study Reported Values, Version 2.2					
Sample ID 100% SSTM					
Sequence	Item	Symbol	Value	Approximately 95% LCL	Approximately 95% UCL
1	Total number of laboratories	p	10		
2	Total number of replicates	Sum(n(L))	120		
3	Overall mean of all data (grand mean)	LPOI or LPOD	0.9667	0.9174	0.9870
4	Repeatability SD	s(r)	0.1784	0.1576	0.2055
5	Among-laboratories SD	s(L)	0.0273	0.0000	0.0930
6	Homogeneity test of laboratory PODs	*P*-value	0.2506		
7	Reproducibility SD	s(R)	0.1804	0.1610	0.2121
8	Intraclass correlation coefficient for repeatability	I(r)	0.9772	0.7818	1.0000

Acknowledgments

We wish to thank the Expert Review Panel for Botanical Identification Methods for kindly reviewing this article and supplying numerous comments for improvement. In particular, we wish to thank Paul Wehling of General Mills/Medallion Laboratories and Danica Reynaud of AuthenTechnologies for the extraordinary amount of time they spent both in reviewing and providing constructive criticism.

References

(1) U.S. Food and Drug Administration (1994) *Dietary Supplement Health and Education Act of 1994,* Washington, DC
(2) U.S. Food and Drug Administration (2007) *Current Good Manufacturing Practice in Manufacturing, Packaging, Labeling, or Holding Operations for Dietary Supplements, Code of Federal Regulations, Title 21, Part III,* U.S. Government Printing Office, Washington, DC
(3) Wehling, P., LaBudde, R.A., Brunelle, S.L., & Nelson, M.T. (2011) *J. AOAC Int.* **94**, 335–347
(4) LaBudde, R.A. (2008) *Statistical Analysis of Interlaboratory Studies, XX, Measuring the Performance of a Qualitative Test Method,* TR290, Least Cost Formulations, Ltd, Virginia Beach, VA
(5) LaBudde, R.A. (2010) *Sampling Plans to Verify the Proportion of an Event Exceeds or Falls Below a Specified Value,* TR308, Least Cost Formulations, Ltd, Virginia Beach, VA
(6) LaBudde, R.A. (2009) *Coverage Accuracy for Binomial Proportion 95% Confidence Intervals for 12 to 100 Replicates,* TR297, Least Cost Formulations, Ltd, Virginia Beach, VA
(7) LaBudde, R.A. (2009) *Statistical Analysis of Interlaboratory Studies, XXII, Statistical Analysis of a Qualitative Collaborative Study as a Quantitative Study Under the Large Sample Approximation,* TR296, Least Cost Formulations, Ltd, Virginia Beach, VA

ANNEX A
SIMCA

Principal component analysis (PCA) is a mathematical procedure used to convert observations for samples with a large number of possibly correlated variables (ions, wavelength, or wavenumbers) into a set of uncorrelated variables called principal components (1). The transformation takes place in manner that assigns the maximum variance to the first principal component with less variance being accounted for by each successive principal component. PCA is applied to the entire data set to determine what groupings of the samples can be seen without any prior decisions (i.e., it is unsupervised). The first two or three principal components (displayed as two- or three-dimensional plots) can be used to demonstrate general patterns in the data.

SIMCA is a supervised approach that builds a PCA model for each specified category of samples (2). Distances between the models are then used to determine the independence of each category of samples. New samples can be assigned to one of the categories or classified as not fitting in any of them.

SIMCA is used for BIMs because predetermined categories of samples are established and modeled. For a BIM, however, only a single PCA model is constructed, and that is for the samples in the inclusivity panel. All other samples are then evaluated using the PCA model to determine whether it is described by the inclusivity PCA model or whether it lies a significant distance from the model, i.e., it does not belong to the inclusivity panel category of samples.

Two statistics used to evaluate whether a sample fits the PCA model are the Q residual and the Hotelling T^2 statistic. The Hotelling T^2 statistic is the multivariate analog of the univariate Students' *t* statistic. It describes how a sample fits in the model. The Q residual, also called the squared prediction error, is more commonly used for process control applications. It describes how far a sample falls outside the model. Some chemometric programs provide both of these statistics as a means of evaluating the fit of a PCA model to the data (1).

Figure A1 provides a simplified illustration of the relationship of the two statistics. In this case, a PCA model is fit to one category of samples. Since only the first principal component was used for this model, the model is a straight line. The data have been mean-centered, so they are centered around the origin, i.e., the intersection of the *x* and *y*-axis. The distribution of each sample with respect to the model is determined by dropping a line from the sample point perpendicular to the model line. The distance from the point where the perpendicular of a sample intersects the model line to the origin provides the Hotelling T^2 value for that point. With sufficient data and a normal distribution, the data distribution should appear as a bell-shaped function centered at the origin. Using this distribution, it can be determined whether a sample is well-fit by the model, i.e., falls inside the 95% confidence limits.

The variance of the sample data with respect to the model is the variance computed along the straight line. In this case, it would be analogous the Students' *t* calculation, i.e., the sum of square of the distance for each sample. In Figure A1, the first principal component for the modeled category passes through the sample data in a manner that provides the maximum variance. A second principal component, perpendicular to the first, would account for the distance of the points from the line and, in this case, provide far less variance than the first principal component. For a model based just on the first principal component, the variance associated with the distance of the sample points from the line is accounted for by the Q residual.

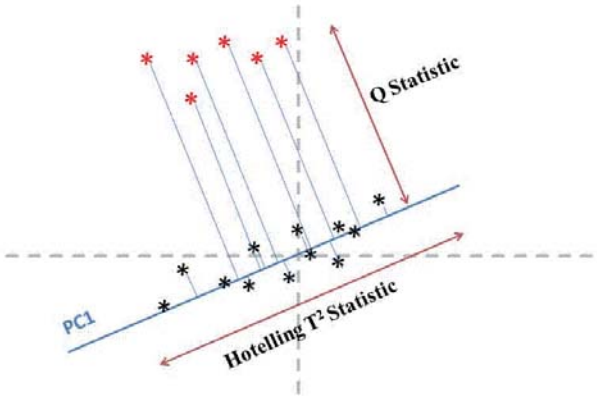


Figure A1. Illustration of Hotelling T^2 and Q statistic: (*) modeled samples and (*) unknown samples.

The distribution of unmodeled data from a second category of samples can be evaluated using the model for the first category of samples. As shown in Figure A1, the distribution of the second category of samples on the first model is very reasonable. Perpendicular lines from the samples in the second category intercept the model line at reasonable distances from the origin. If this were real data, and a 95% confidence limit had been computed, the second category of samples would undoubtedly be within that limit. However, for the second category of samples, a much larger fraction of the total variance is incorporated in the distance from the model line. The second category samples will fall well outside the 95% confidence limit for the Q residual established by the first category samples.

SIMCA can be applied to a BIM by constructing a PCA model using the data from the inclusivity panel botanical materials. New samples are fit to the model and the Q residual is determined. If the Q residual for a sample falls outside the 95% confidence limit, the new sample is not the same as the target materials. Conversely, if the new sample falls within the 95% confidence limit, it would be classified as a target material.

References

(1) Wold, S., & Sjostrom, M. (1977) in *Chemometrics Theory and Application*, American Chemical Society Symposium Series 52, American Chemical Society, Washington, DC, pp 243–282
(2) Wold, S. (1987) *Chemom. Intel. Sys*. **2**, 37–52

ANNEX B
Modeling of the POI Using Logistic Regression

The models in common use for this kind of problem include, among many others: (*1*) discriminant analysis; (*2*) logistic regression; or (*3*) normit regression. There is also a choice of metamer x (i.e., transform of %SSTM). Common choices include x = % SSTM, or x = $\log_{10}$ (%SSTM + 0.5). Logistic and normit regression assume the POI versus x curve is symmetrical, which that of Figure 4 obviously is not.

Suppose we choose logistic regression with an identity metamer (x = % SSTM), which implies the model:

```
Call:
glm(formula = cbind(id, notid) ~ x, family =
binomial("logit"),
    data = dat)
Deviance Residuals:
      1       2       3       4
 0.8314  0.9386 -1.5687  2.6222
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.04711    0.67021  -7.531 5.05e-14 ***
x            0.07878    0.01001   7.869 3.57e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
(Dispersion parameter for binomial family taken to be
1)
    Null deviance: 186.241  on 3  degrees of freedom
Residual deviance:  10.908  on 2  degrees of freedom
AIC: 25.12
Number of Fisher Scoring iterations: 5
```

Figure B1. Fit of Equation 1 to the sample data.

$$\text{logit(POI)} = \ln\{\text{POI}/(1 - \text{POI})\} = \alpha + \beta x = \alpha + \beta\ (\%\ \text{SSTM}) \quad \text{(Equation 1)}$$

For the sample data, the fit is as shown in Figure B1.

The model fits poorly and is highly overdispersed (dispersion = 10.908 / 2 = 5.454). Consequently, the standard errors found in the fit should be multiplied by 2.34 = √5.454. (Note that this overdispersion suggests that the logistic regression model with specified link is a poor choice for the data.)

An estimate of the point at which POI = 0.5000 is given by the negative ratio of the intercept by the slope, or x = 64.1% SSTM. This would be denoted "Effective Concentration at POI = 0.50" or "EC50." (It should be noted that EC50 depends upon the definitions of the SSTM and SITM.)

From the logistic regression fit, we get the results shown in Table B1 and Figure B2. The logistic regression does not do as well as the direct POI descriptive statistics of Table 6, because of serious failure of the model assumptions. (It turns out that *none* of the usual generalized model forms fits the asymmetrical POI versus % SSTM curve very well for this example. So it should be noted that the standard error of POI is *not* always reduced by fitting across the combination of concentrations used.) Note that, based on the logistic model, the BIM continues to pass the 0% SSTM performance requirement, but fails the 100% SSTM requirement.

It is generally recommended that the methods of Table 6 be used for evaluating performance requirements, rather than those of unvalidated regression models. One of the advantages, however, of fitting such a model is that continuous curves may be obtained, as shown in Figure B3.

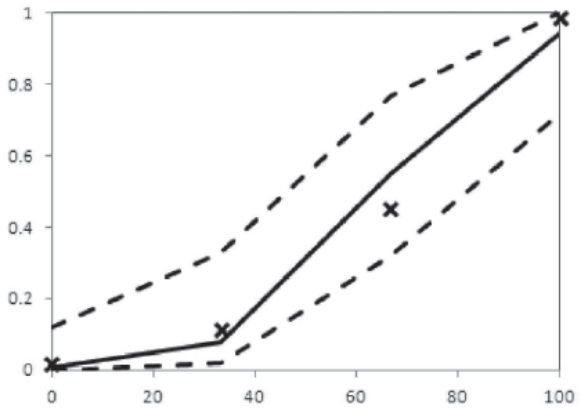


Figure B2. Example SLV results from a logistic regression fit showing POI (solid line), lower 95% confidence limit (dashed line below the POI), and upper 95% confidence limit (dashed line above the POI), and measured POI values (X).

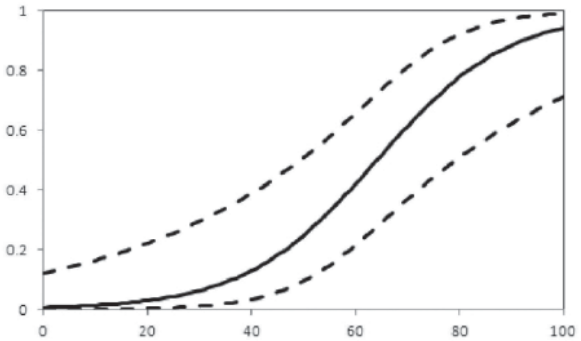


Figure B3. Continuous curves from SLV logistic regression fit showing POI (solid line), lower 95% confidence limit (dashed line below the POI), and upper 95% confidence limit (dashed line above the POI).

Table B1. SLV results (logistic regression fit)

	Fitted	Obs.	1-sided	LCL	UCL
% SSTM	POI	POI	95%	95%	95%
0.0	0.0064	0.0167	0.0778	0.0003	0.1214
33.3	0.0816	0.1167		0.0162	0.3239
66.7	0.5511	0.4500		0.3181	0.7636
100.0	0.9443	1.0000	0.7715	0.7126	0.9915