Probability of Identification (POI): a Statistical Model for the Validation of Qualitative Botanical Identification Methods

Robert A. LaBudde
Least Cost Formulations, Ltd., 824 Timberlake Drive, Virginia Beach, VA 23464
Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529

James M. Harnly[1]
Food Composition and Methods Development Lab, Beltsville Human Nutrition Research Center, Agricultural Research Center, US Department of Agriculture, Building 161 BARC-East
Beltsville, MD 20705

Abstract

A qualitative botanical identification method (BIM) is an analytical procedure which 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 non-target (undesirable) material. We describe the development and validation of studies for a BIM based on the idea of a 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 multi-collaborative study examples are given.

[1]Corresponding author at email: james.harnly@ars.usda.gov

1. Introduction

A "Botanical" is an herbal type of dietary supplement regulated in the USA 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 (cGMPs) for foods and dietary supplements (2) issued by FDA has tasked manufacturers with authentication of 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 which 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 Wehling, 2011) shows a plot where the concentration of the target material increases towards the right while the concentration of a non-target 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 symmetric, POI plots are usually asymmetric. The POI model is based on the probability of detection (POD) model which was developed for binary qualitative methods for detection of biological threat agents (3,4).

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 non-target material increases. The goal of method development and validation is to first determine if the method meets standard method performance requirements (SMPRs), and secondly to characterize how the method makes the transition from a negative to a positive response.

The SMPRs will specify the target botanical materials (inclusivity sampling frame), the non-target materials (exclusivity sampling frame), the physical form of the materials, the minimum concentration of target material that is acceptable in the presence of non-target 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 case 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.

2. 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 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.

Botanical Identification Method (BIM): A method that establishes identity specifications for a botanical material and allows determination, within a specified statistical limit, that a test material is a true example of the target botanical material and meets the identity specifications. Thus, a method answers the question, "*Is the test material the same as the target material?*" not "*What is this material?*" In most cases, the method will be validated to 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/non-consistent) answer.

Candidate Method: The method to be validated.

Exclusivity: Ability of a BIM to correctly reject non-target botanical materials.

Exclusivity Sampling Frame (ESF): A list of practically obtainable non-target 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 exclusivity sampling frame 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 methods (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.

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 sampling frame should be sufficiently large that the botanical variation is adequately represented. Sources of variation may include, but are not limited to, species, sub-species, cultivar, growing location, growing conditions, growing season, and post-harvest processing.

Inclusivity Panel: A subset of the inclusivity sampling frame 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.

Non-Target 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 SMPRs.

Probability of Identification (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.

Specified Inferior Test Material (SITM): A mixture of botanical materials that contains 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., identify 0% of the time). The SITM will typically be high quality target material mixed with the worst-case (for identification) non-target material.

Specified Superior Test Material (SSTM). A mixture of botanical material that contains 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 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) non-target material.

Standard Method Performance Requirements (SMPRs): Performance requirements based on the fitness for purpose statement for each method. For BIMs, the SMPRs should minimally include the physical form of the sample, the ISF, the ESF, the SSTM, and the SITM.

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.

3. 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 sub-species 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 8 international locations (80 samples). In reality, only 25 of the desired samples may be practically obtainable. These 25 obtainable samples comprise the inclusivity sampling frame (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 SMPRs 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 the variant be tested. Age of the plant may be a factor of importance. The subset of this list which is practically obtainable for a validation study is the ISF.

The SMPRs 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 SMPRs 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.

4. Exclusivity Panel

The list of non-target 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 exclusivity sampling frame (ESF) will be comprised of those botanical materials that are practically obtainable. The exclusivity panel will comprise those samples used for method development and validation.

The SMPRs must provide a list of all necessary or commonly encountered non-target 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 the non-target material be tested. The subset of this list which is practically obtainable for a validation study should then be identified as the ESF.

The SMPRs should identify the minimum number of non-target 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 to comprise a set as specified by the ERP. More replicates and randomization may allow a quantitative statistical inference to be made concerning exclusivity.

5. 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 non-target materials listed in the ESF. The BIM should clearly and unequivocally discriminate between the target and non-target 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 SMPRs.

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 SMPRs) to characterize the POI on the variant quantitatively. If the variant fails to meet minimum acceptable performance requirements as set by the SMPRs, 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 is the actual results obtained, but without suggestive quantitative statistics. For example, without randomization, the use of percentages or other quantitative measures is inappropriate.

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

After inclusivity and exclusivity studies have been completed, target and non-target material(s) are chosen to verify that the method can discriminate between the SSTM and the SITM. Either the worst-case non-target materials or perhaps the most common non-target materials would typically be chosen. In addition, a combination of target and non-target 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 SMPRs.

The SMPRs 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 non-target material(s).

7. 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 univariate analytical parameter (AP) that will be used to determine whether the test sample matches 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 is usually based on the characterization of the POI as a function of target material concentration.

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 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 of a single sample are analyzed, then the width of the distribution will be narrow (a smaller standard deviation) 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 SMPRs. 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 was 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 must be prepared that fall between the SSTM and SITM. In each case, the threshold will intersect each peak and determine the POI. As the SSTM:SITM values change from1:0 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 the shape of the distribution functions involved.

8. A Specific Example – American ginseng mixed with Chinese ginseng

A data set is presented here that illustrates the analytical measurements discussed in the previous section. The target botanical material is American ginseng and the non-target material is Chinese ginseng. The inclusivity panel consists of 43 American ginseng (AG) samples grown in the US (harvested over 3 years from 20 different farms in Wisconsin) and the exclusivity panel consists of 8 Chinese ginseng (CG) samples grown in China (Table 1).

The AG and CG samples were analyzed by direct injection mass spectrometry 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 x 8 CG).

The multivariate data (1000 variables) were analyzed using SIMCA (soft independent modeling of class analogy) (Appendix A). SIMCA produced a single univariate value, the Q residual, that 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 (Appendix A).

Figure 4A shows the inclusivity/exclusivity study. The Q residual is plotted for individual samples. With 100% AG (inclusivity panel samples) as the model, it can be seen that 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 that reflects the 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 SMPRs for this model. Figure 4B shows the SSTM samples (98% AG) as well as 100% AG and 100% CG samples. The pattern of 8 groupings for the SSTM samples reflects that all 43 AG samples were diluted by each of the 8 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 4C 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 5A. 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. It can be seen that if the SSTM were specified at a lower concentration of AG or the SITM at a higher concentration of AG, the method would not be appropriate unless lower confidence limits were chosen.

Based on the AP threshold shown in Figures 4B, 4C, and 5, the POI in Figure 5B 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 with 95% confidence. Table 2 shows that, for these performance requirements, 60 replicates must be tested at each level with no more than 2 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 may be selected from Table 3. For more plans, see LaBudde (7).

9. 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, so 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 1= Identified or 0 = Not Identified is returned. If there are no SMPRs, the method would be "characterized" as opposed to "validated".

Consider testing in a single independent laboratory, or a "single laboratory validation" (SLV). With respect to the performance requirements of Table 2, the SITM and SSTM were used to prepare mixtures in the proportions 0%:100%, 33%:67%, 67%:33% and 100%:0%. From each of these mixtures, 60 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 false positive fraction (FPF) performance requirement succeeds at

0% SSTM, because no more than 2 test portions reported identification. Also, the false negative fraction (FNF) performance requirement at 100% SSTM succeeds because, in both cases, fewer than 2 test portions were not identified.

Using the methods of Wehling(3) and LaBudde(5,6), 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 (viz., % SSTM) are known with certainty here, one of several regression models might be fit to possibly (although this is not guaranteed) obtain more precise estimates of POI and its confidence limits, but at the expense of some additional assumptions (see Appendix B).

10. 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 6 collaborators does not allow inference with respect to the general collaborator population, 8 collaborators allows reasonable estimation, and 10 collaborators is desirable. More than 10 collaborators is useful, but not necessary. For fewer than 6 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 is therefore 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 less than 6 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.

Concentration levels (i.e., percentage of SSTM in a SSTM:SITM mix) 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 necessary. Therefore the smallest acceptable collaborative study final data would be 6 collaborators x 12 replicates = 72 test portions.

It should be noted that a performance requirement imposed on a collaborative study will, due to the inter-collaborator variation, be more difficult for a candidate BIM to achieve than would be a performance requirement imposed on a single laboratory validation 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 and 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 of replicates (see Section 9). 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.

Now 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 1 or 2 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%, 100%).

After completion of the study, two collaborators were removed for cause, and the results shown in Table 6 were 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, and the 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 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 givesthe results in Table 9. Once again is no detected intercollaborator effect (p-value = 0.18), so there is little evidence 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, and the 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 the method is irreproducible, and the method meets the POI (or FNF) performance requirement.

11. Lot-Lot Variability Study, Time Stability Study and Robustness Study

The single laboratory 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. Just as with collaborators in a collaborative study, estimation of the lot random effect requires at least 6 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 there is no material degradation in performance over the life of lots of materials and equipment. This may be accomplished by determination of 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.

12. Conclusion

The purpose of a qualitative BIM is to discriminate between acceptable target material and target material with an unacceptable concentration of non-target 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 has been given which allows validation and/or characterization of qualitative BIMs. Examples have been given

for both single laboratory and collaborative studies with SMPRs. 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.

13. Acknowledgements

The authors 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 Medallion Laboratories and Danica Reynaud of AuthenTechnologies for the extraordinary amount of time they spent in both review and providing constructive criticism.

14. References

1. Food and Drug Administration of the United States (1994), Dietary Supplement Health and Education Act of 1994 (DSHEA).

2. Food and Drug Administration of the United States (2007), 21 CFR part 111, Current Good Manufacturing Practice (CGMP) In Manufacturing, Packaging, Labeling, Or Holding Operations For Dietary Supplements (72 FR 34752).

3. Wehling, P, LaBudde R.A., Brunelle S.L., Nelson M.T. (2011). "Probability of detection (POD) as a statistical model for the validation of qualitative methods." J. AOAC 94(1): 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. (2009). "Coverage accuracy for binomial proportion 95% confidence intervals for 12 to 100 replicates." TR297. Least Cost Formulations, Ltd., Virginia Beach, VA.

6. 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.

7. LaBudde, R.A. (2010b). "Sampling plans to verify the proportion of an event exceeds or falls below a specified value." TR308. Least Cost Formulations, Ltd., Virginia Beach, VA.

Table 1: *Panax* samples analyzed in this Study

N	Label	Provider	Source
Inclusivity Panel (American ginseng)			
26	American ginseng	Ginseng Board of Wisconsin	USA
13	American ginseng	American Herbal Pharmacopoeia	USA
4	American ginseng	Internet Retailer (Wisconsin farm)	USA
ExclusivityPanel (Chinese ginseng)			
3	Chinese ginseng, red	American Herbal Pharmacopoeia[2]	China
1	Kirin Red #1	Internet Retailer	China
1	Kirin Red #3	Internet Retailer	China
1	Kirin Red #5	Internet Retailer	China
1	Shih Chu #25	Internet Retaile	China
1	Shih Chu #80	Internet Retailer	China
SSTM*			
344	0.98 American ginseng + 0.02 Chinese ginseng		
SITM*			
344	0.90 American ginseng + 0.10 Chinese ginseng		

*In each case, each of the 43 American ginseng samples were mixed with each of the 8 Chinese ginseng samples (43x8=344)

Table 2. Example performance requirements

Requirement	***% SSTM***	***Measure***	***Maximum***	***# Replicates to be tested***	***# Failures allowed***
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
NOTE: In each case, no more than 2 failures are allowed.					

623

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]	***# Replicates to be tested***	***# 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(7).						
[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.						

624
625

Table 4. Observed SLV Results for Example BIM				
% SSTM	***# Test Portions***	***# Identified***	***# 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

626
627

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

628
629

630 Table 6: Collaborative study results.

% SSTM	Collaborator	Replicates	# 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

631
632
633

Table 7: Collaborative study results for 0 % SSTM concentration.

AOAC Binary Data Interlaboratory Study Workbook ***Version:*** ***2.2***

Study Reported Values

Seq.	Item	Symbol	Value	Appox. 95% LCL	Appox. 95% UCL
	Sample ID:		0 % SSTM		
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 standard deviation	s(r)	0.0913	0.0807	0.1713
5	Among-Laboratories standard deviation	s(L)	0.0000	0.0000	0.0402
6	Homogeneity test of laboratory PODs	p-value	0.4303		
7	Reproducibility standard deviation	s(R)	0.0913	0.0814	0.1064
8	Intraclass correlation coefficient for repeatability	I(r)	1.0000	0.8335	1.0000

Table 8. Collaborative study results for 33.33 % SSTM concentration.

AOAC Binary Data Interlaboratory Study Workbook ***Version:*** ***2.2***

Study Reported Values

Seq.	Item	Symbol	Value	Appox. 95% LCL	Appox. 95% UCL
	Sample ID:		33.33 % SSTM		
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 standard deviation	s(r)	0.3703	0.3272	0.4266
5	Among-Laboratories standard deviation	s(L)	0.0000	0.0000	0.1400
6	Homogeneity test of laboratory PODs	p-value	0.6563		
7	Reproducibility standard deviation	s(R)	0.3703	0.3304	0.4275
8	Intraclass correlation coefficient for repeatability	I(r)	1.0000	0.8889	1.0000

647 Table 9: Collaborative study results for 66.67 % SSTM concentration.
648

AOAC Binary Data Interlaboratory Study Workbook ***Version:*** ***2.2***

Study Reported Values

Seq.	Item	Symbol	Value	Appox. 95% LCL	Appox. 95% UCL
	Sample ID:		66.67 % SSTM		
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 standard deviation	s(r)	0.4939	0.4364	0.5222
5	Among-Laboratories standard deviation	s(L)	0.0948	0.0000	0.2779
6	Homogeneity test of laboratory PODs	p-value	0.1783		
7	Reproducibility standard deviation	s(R)	0.5029	0.4489	0.5222
8	Intraclass correlation coefficient for repeatability	I(r)	0.9644	0.7547	1.0000

649
650
651
652
653 Table 10: Collaborative study results for 100.0 % SSTM concentration.
654

AOAC Binary Data Interlaboratory Study Workbook ***Version:*** ***2.2***

Study Reported Values

Seq.	Item	Symbol	Value	Appox. 95% LCL	Appox. 95% UCL
	Sample ID:		100.0 % SSTM		
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 standard deviation	s(r)	0.1784	0.1576	0.2055
5	Among-Laboratories standard deviation	s(L)	0.0273	0.0000	0.0930
6	Homogeneity test of laboratory PODs	p-value	0.2506		
7	Reproducibility standard deviation	s(R)	0.1804	0.1610	0.2121
8	Intraclass correlation coefficient for repeatability	I(r)	0.9772	0.7818	1.0000

655
656
657
658

Figure Captions

Figure 1 Probability of Identification for botanical identification.

Figure 2 Inclusivity/exclusivity and SSTM/SITM characterization.

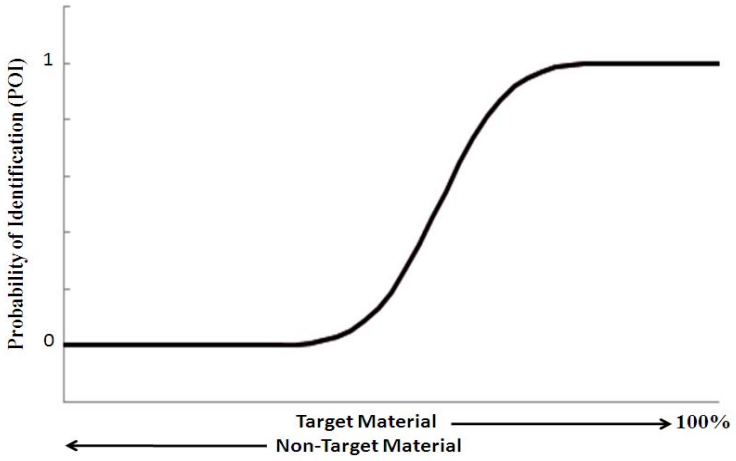
Figure 3 Conversion of SSTM, SITM, and intermediate concentrations to POI.

Figure 4 SIMCA plots for A) 100% American ginseng and 100% Chinese ginseng, B) SSTM, 100% American ginseng, and 100% Chinese ginseng, and C) SITM, 100% American ginseng, and 100% Chinese ginseng. Legend: **Red** – American ginseng (AG), **Green** – Chinese ginseng (CG), and **Blue** – B) SSTM and C) SITM.

Figure 5 Target material American ginseng, non-target material Chinese ginseng: A) SITM and SSTM and B) POI.

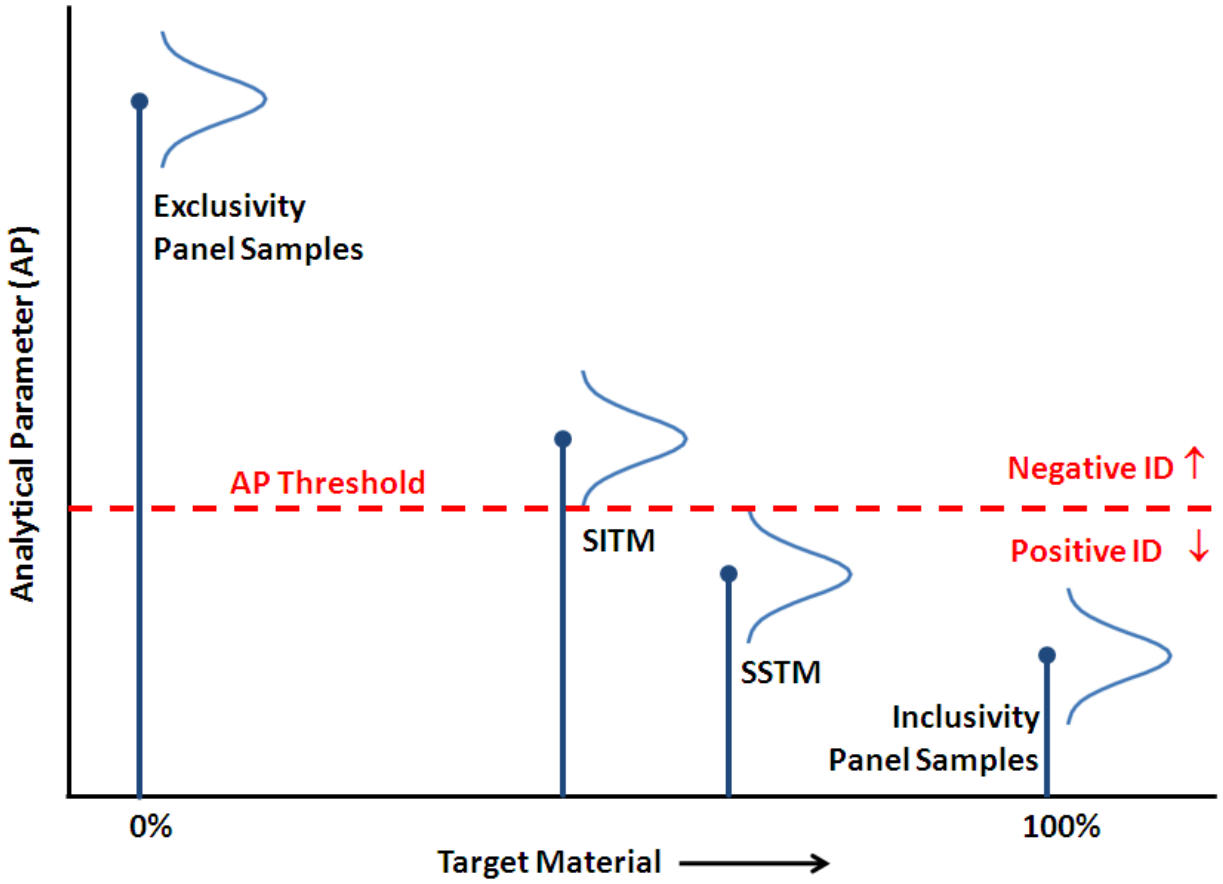
Figure 6 Expected POI vs. % SSTM for an example BIM. Note the POI at 0% is the false positive fraction. 1- POI at 100% is the false negative fraction.

679 Figure 1: Probability of Identification for botanical identification
680
681 **A**



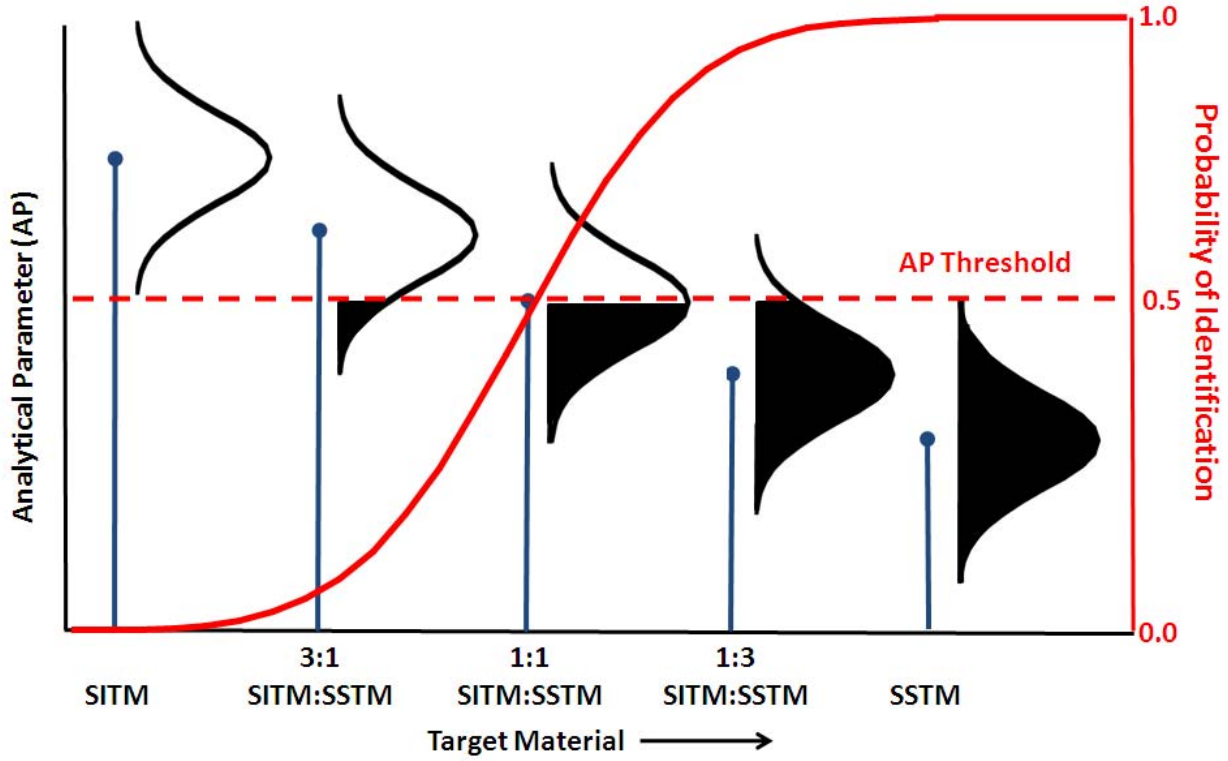
682
683
684
685

686 Figure 2: Inclusivity/exclusivity and SSTM/SITM characterization.
687
688



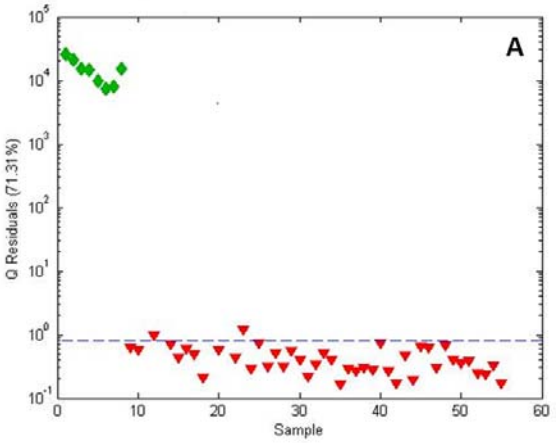
689
690

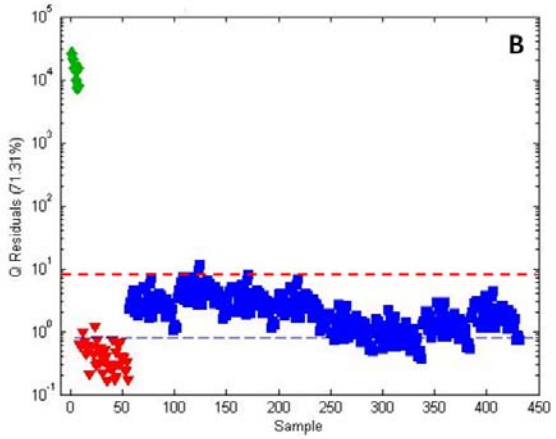
691 Figure 3: Conversion of SSTM, SITM, and intermediate concentrations to POI.
692
693



694
695

Figure 4: SIMCA plots for A) 100% American ginseng and 100% Chinese ginseng, B) SSTM, 100% American ginseng, and 100% Chinese ginseng, and C) SITM, 100% American ginseng, and 100% Chinese ginseng. Legend: **Red** – American ginseng (AG), **Green** – Chinese ginseng (CG), and **Blue** – B) SSTM and C) SITM.





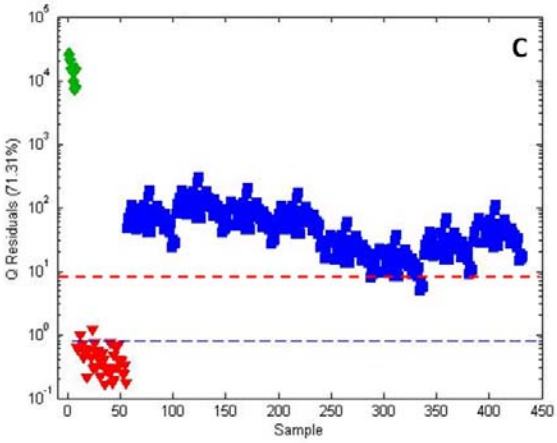
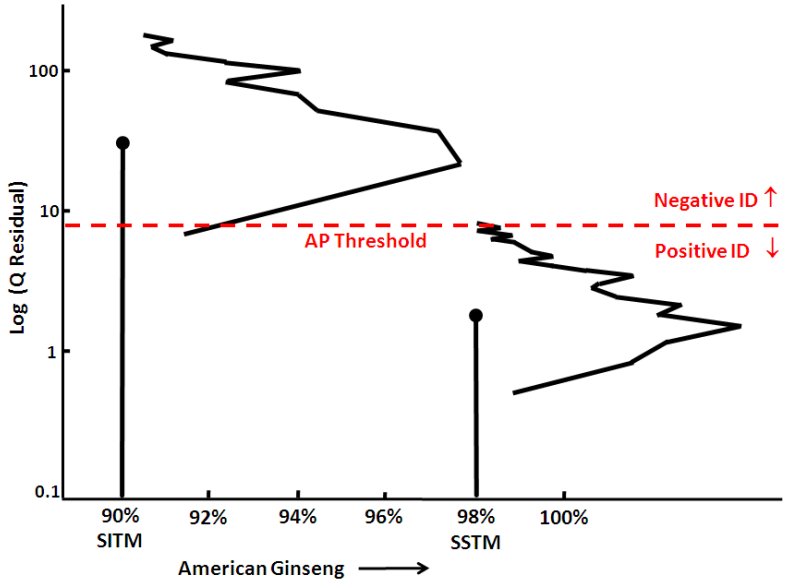
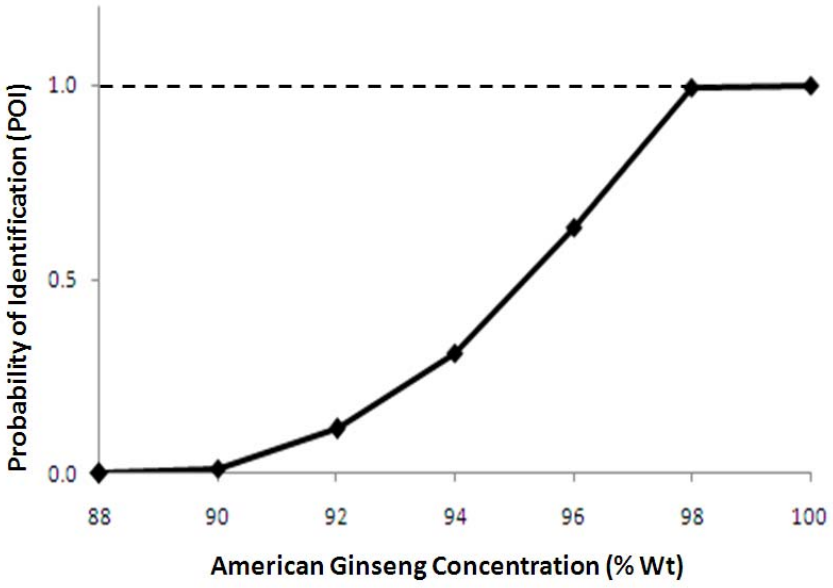


Figure 5: Target material American ginseng, non-target material Chinese ginseng: A) SITM and SSTM and B) POI.

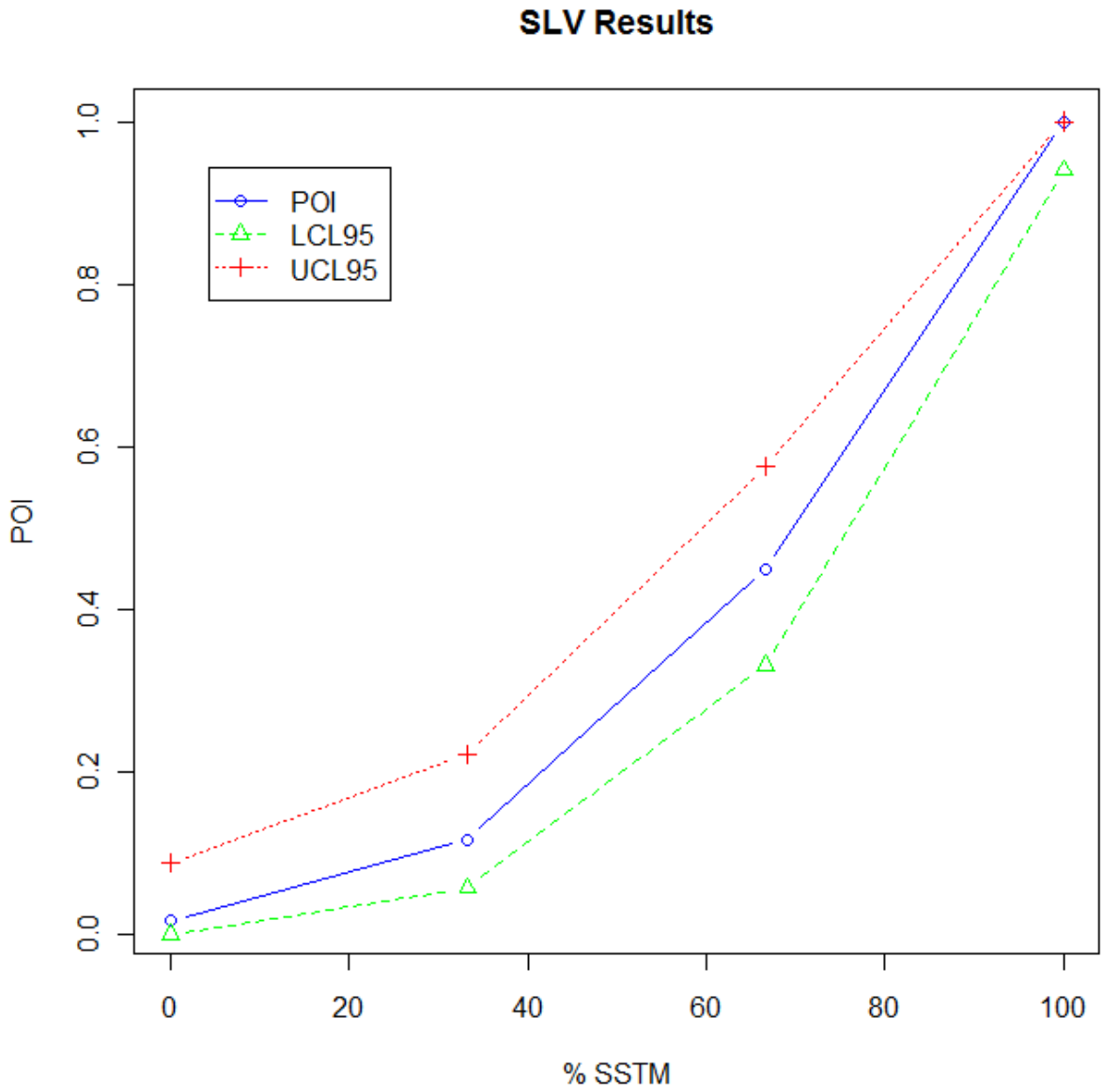
A



B



716 Figure 6: Expected POI vs. % SSTM for an example BIM. Note the POI at 0% is the false
717 positive fraction. 1- POI at 100% is the false negative fraction.



718
719

Appendix A: SIMCA

Principal component analysis (PCA) is a mathematical procedure that is 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 (A1). 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 2 or 3 principal components (displayed as 2 or 3 dimensional plots) can be used to demonstrate general patterns in the data.

SIMCA (soft independent modeling of class analogy) is a supervised approach that builds a PCA model for each specified category of samples (A2). 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 doesn't belong to the inclusivity panel category of samples.

Two statistics that are 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 (SPE), 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 (black dots). 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.

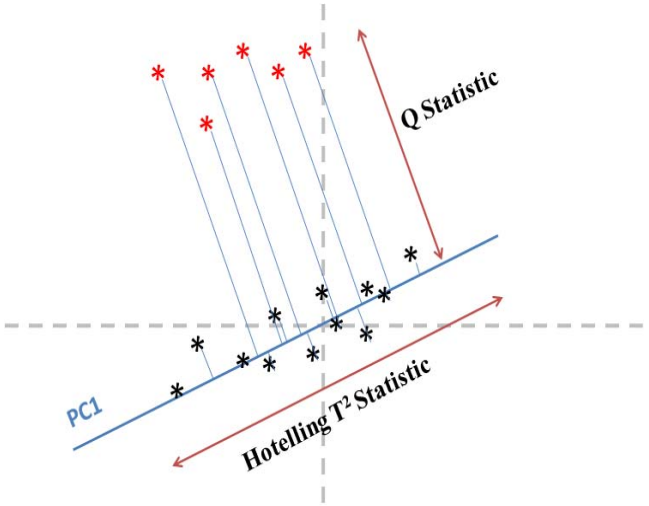
The variance of the sample data with respect to the model is the variance computed along the straight line. In this case it would be analogous the Students' t calculation, i.e. the sum of square of the distance for each sample. In Figure A1, the first principal component for the modeled category (black dots) 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.

The distribution of un-modeled data from a second category of samples (the red dots) 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. 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.

Figure A1: Illustration of Hotelling T2 and Q statistic: ∗) modeled samples and ∗) unknown samples.



References:

A1 Wold, Svante, and Sjostrom, Michael (1977) SIMCA: A method for analyzing chemical data in terms of similarity and analogy, in Kowalski, B.R., ed., Chemometrics Theory and Application, American Chemical Society Symposium Series 52, Wash., D.C., American Chemical Society, p. 243-282.

A2 Wold, S. (1987) Principal component analysis. *Chemom. Intel. Sys*. 2, 37-52.

Appendix 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 vs. x curve is symmetric, which that of Fig. 4 obviously is not.

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

$$\text{logit(POI)} = \ln\{\text{POI} / (1 - \text{POI})\} = \alpha + \beta\, x = \alpha + \beta\ (\%\ \text{SSTM}) \qquad \text{(C.1)}$$

For the sample data, the fit is:

```
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
```

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 these 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.)

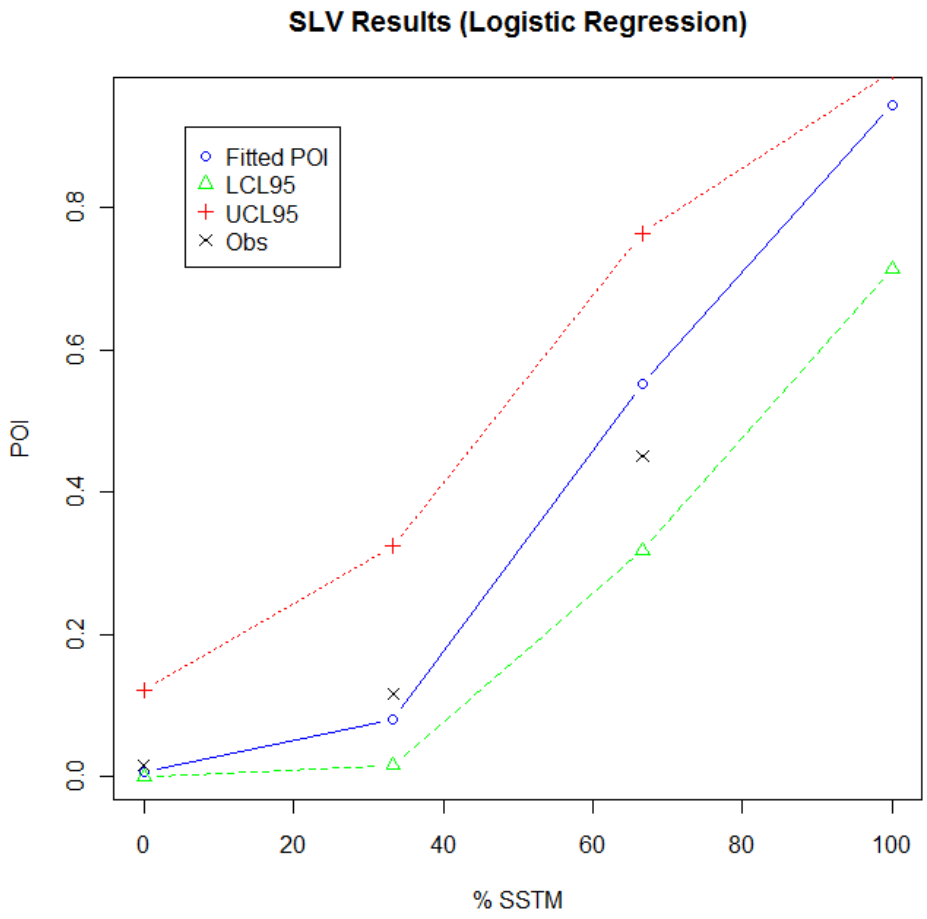
From the logistic regression fit, we get the results shown in Table B1 and Figure B1. The logistic regression does not do as well as the direct POI descriptive statistics of Table 6, because of serious failure of the model assumptions. (It turns out that *none* of the usual generalized model forms fits the asymmetric POI vs. % 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 B2.

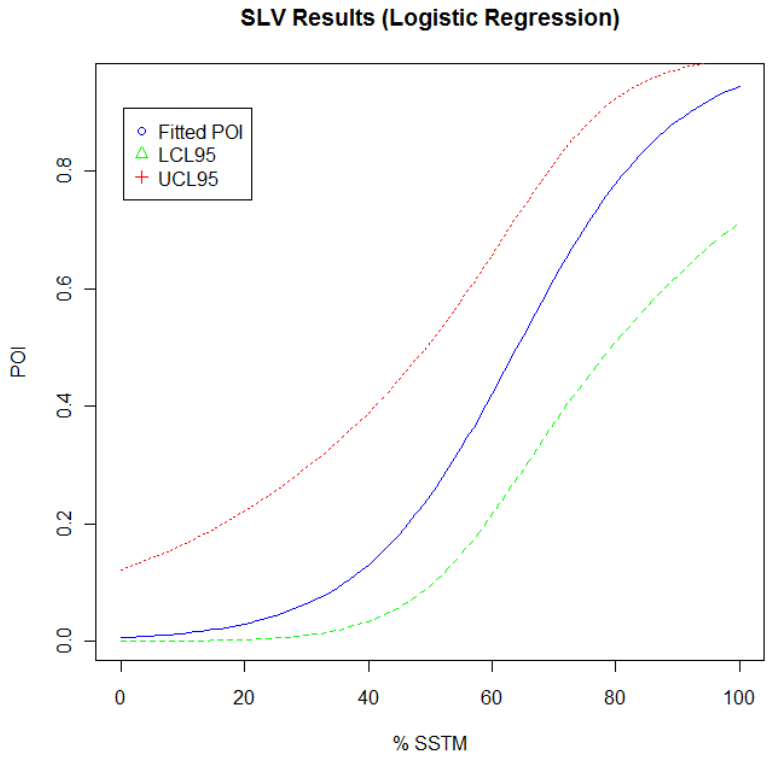
Table B1. SLV Results (Logistic Regression Fit)

% SSTM	Fitted POI	Obs. POI	1-sided 95%	LCL 95%	UCL 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

Figure B1: Example SLV results from a logistic regression fit.



866 Figure B2: Continuous curves from SLV logistic regression fit.



867