Appendix N: ISPAM Guidelines for Validation of Qualitative Binary Chemistry Methods

1.0 Scope

The purpose of this document is to provide a guideline for the validation of binary [a binary qualitative method is one that produces one out of two possible responses when it is used (e.g., end time PCR, visual inspection of a dip stick)] qualitative methods intended to detect biological and chemical compounds. Qualitative methods that are used to make a detection decision by comparing the value of a response to a cut-off value should be validated by using quantitative statistics, where possible, on the responses (the use of quantitative statistics usually gives better estimates of method performance for fewer replicate analyses in each laboratory).

2.0 Terms and Definitions

Where appropriate, definitions have been taken from international standards and the source is noted. Sources of definitions include the following:

ISO/IEC Guide 99:2007, *International vocabulary of metrology—Basic and general concepts and associated terms (VIM)*

ISO 3534-2:2006, *Statistics—Vocabulary and symbols—Part 2: Applied statistics*

ISO 14971:2007, *Medical devices—Application of risk management to medical devices*

ISO 17511:2003, *In vitro diagnostic medical devices—Measurement of quantities in biological samples—Metrological traceability of values assigned to calibrators and control materials*

ISO 5725-1:1994, *Accuracy (trueness and precision) of measurement methods and results—Part 1: General principles and definitions*

USP 31:2008, *U.S. Pharmacopeia General Information/<1223> Validation of Alternative Microbiological Methods*

Candidate method.—The method submitted for validation.

Matrix.—Totality of components of a material system except the analyte (ISO 17511).

Method.—A procedure that includes sample processing, assay, and data interpretation.

Probability of detection (POD).—The proportion of positive analytical outcomes for a qualitative method for a given matrix at a given analyte level or concentration. POD is concentration dependent.

Qualitative binary method.—A method of analysis with two possible outcomes.

Reproducibility.—Precision under reproducibility conditions (ISO 5725-1).

Reproducibility conditions.—Conditions where test results are obtained with the same method on identical test material in different laboratories with different operators using different equipment.

Sample.—A small portion or quantity taken from a population or lot that is ideally a representative selection of the whole. Samples are taken from lots for purposes of scientific examination and analysis and are intended to provide characteristic information about the population, generally by applying statistical calculations. Source: ISO 3534-1:1993.

Laboratory sample.—Sample as prepared for sending to the laboratory and intended for inspection or testing. Source: ISO 7002:1986.

Test portion.—A fraction of a sample intended for analysis. There are cases (liquid products, analysis of symptoms, etc.) where the laboratory sample is also the test sample. Source: ISO 21572.

3.0 Selectivity Study

The selectivity study is a single-laboratory study designed to demonstrate that a method does not detect non-target compounds, and at the same time demonstrate a candidate method's ability to detect the related compounds.

3.1

Organize a "selectivity" test panel of related compounds that are expected to give a positive result. Document the source and origin of each test panel compound. All documentation of the analyte identity must be on file and available for review.

3.2

Organize a panel of non-target compounds that might be expected to be encountered when the method is used; or to be erroneously detected by virtue of chemical or other similarities.

3.3

Prepare at least one replicate of each target compound from the selectivity test panel at the 95% POD concentration. Prepare at least one replicate of each non-target compound from the selectivity test

The core group that was instrumental in the development of these guidelines consisted of Bert Pöpping, Eurofins CTC GmbH, Hamburg, Germany, and chair of the working group on the development of the guidelines; Joe Boison, Canadian Food Inspection Agency, Saskatoon, Saskatchewan, Canada; Scott Coates, AOAC INTERNATIONAL, Gaithersburg, MD, USA; Christoph von Holst, European Commission, DG Joint Research Centre, Geel, Belgium; Roy Macarthur, The Food and Environment Research Agency, York, UK; Ray Shillito, Bayer CropScience LP, Morrisville, NC, USA; and Paul Wehling, General Mills, Inc., Minneapolis, MN, USA.

The guidelines were reviewed by the AOAC International Stakeholder Panel on Alternative Methods (ISPAM) and approved on March 14, 2013.

panel at an appropriate concentration. Blind code and randomly mix the selectivity and non-target compounds.

3.4

An analyst (or analysts) not involved in the preparation of the test panel shall evaluate the compounds using the candidate method and record the results.

3.5

If an individual test panel compound yields an incorrect result (a negative in the case of a target compound; a positive in the case of a non-target compound) then the compound may be retested with a number of replicates to be determined by subject matter experts. The number of replicates will determine the lower confidence interval for the POD estimate.

4.0 Matrix and POD Concentration Study

The matrix study is a single-laboratory study designed to demonstrate that a candidate method can detect the target compound in the claimed matrixes. Analyze test portions of the claimed matrices containing the target compound(s) at various concentrations. The number of different matrices to be tested depends on the claims and intended use of the method.

In general, a minimum of five concentrations per target compound should be evaluated for each matrix, but more concentrations could be included at the discretion of subject matter experts.

The number of replicates at each concentration should be determined by the subject matter expert(s). The number of replicates at the 95% POD concentration may be greater than at other concentrations. For example, the 95% POD concentration may have 96 replicates while other concentrations only four. A more balanced approach would spread replicates across all concentrations, with a minimum of 20 replicates at each of five concentrations. Some discretion is allowed with consultation by a statistician. For example, if more than five concentrations are desired, the number of replicates per concentration could be reduced. The decisions on number of replicates should be made with an understanding of the desired level of confidence in the final results.

4.1 Incurred or Fortified

A target compound in a matrix may be incurred or fortified. Incurred target compound(s) are preferred. If not available, matrix fortified with the target compound(s) may be used. If a matrix with incurred target compound(s) is used, then matrix that is known to be free of the target compound(s) can be used to 'dilute' it to the desired concentration. Evidence supporting homogeneity must be provided.

4.2 Raw and/or Processed Materials

Both processed (e.g., such as cooked, fermented, etc.) and raw samples should be represented if the assay claims to detect the target compound(s) in such foods.

If the method detects more than one target compound simultaneously in the same test portion, the study should be designed so that the target compounds are fortified together into some of the test portions.

Collect enough of each matrix to prepare more than the required number of test portions for each concentration.

Prepare the required number of test portions of the matrix with the target compound(s) at the specified concentration. Blind-code, randomize, and analyze the prepared test portions.

The analyst performing the analyses should not have knowledge of the study design or the blind codes of the test portions. The analyst should be informed that the design of the study does include a certain number of "blank" samples and that both positive and negative outcomes should be expected.

Plot the response of the method as POD response vs. concentration of target compound(s).

4.3 Statistical Analysis

Refer to *Annex A* for guidance on statistical analysis of data.

5.0 Collaborative Study

A collaborative study characterizes the performance parameters (e.g., POD, repeatability, reproducibility) of the candidate method across testing sites.

Methods shall be validated under conditions of intended use. For example, a method intended for use by trained factory operators at a grain inspection site must be validated under conditions that simulate the grain inspection site and should include representative end users as collaborators.

A collaborative study must include a minimum of 10 testing sites, each reporting at least six valid replicate analyses per concentration. *See Annex B* for recommendations on the range and number of concentrations, and the number of replicates for each concentration. Deviations from these recommendations must be documented and justified.

Study test portions must be blind-coded and shipped to each collaborator. Collaborators shall perform all analyses independently.

If the method detects more than one target compound simultaneously in the same test portion, the study should be designed so that the target compounds are fortified together into some of the test portions.

5.1 Data Analysis

5.1.1 Raw Data Tables

Each test site must report the results obtained with each test portion plus any comments.

5.1.2 Statistical Analysis

Refer to *Annex A* for guidance on statistical analysis of data.

5.1.3 Collaborator Comments

Comments on the candidate method should be collected from all collaborators and reported in the collaborative study report.

ANNEX A
Validation for Binary Qualitative Methods of Detection

Binary qualitative methods are those that give two responses which can usually be interpreted as "target compound(s) detected" or "target compound(s) not detected." Their performance can be validated by collaborative trials, single-laboratory validation studies, or by using observed long-term performance in much the same way, practically, as analytical methods that give quantitative responses. As for quantitative methods, proper consideration must be given to ensuring that the data used to validate methods covers a representative scope and range with adequate replication between laboratories, analysts, or days. The chief practical difference is that more analytical replicates are needed within each analytical condition (laboratory, day, concentration, etc.) when validating qualitative methods.

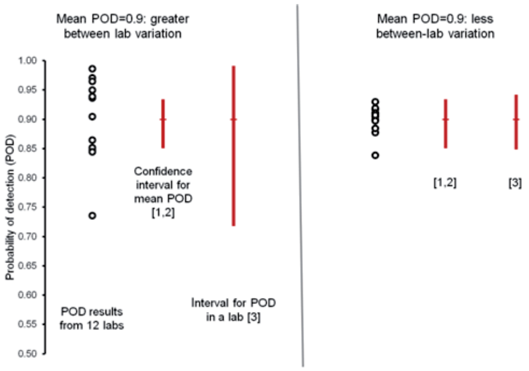

Figure A1. Intervals that describe different aspects of method performance: the average probability of detection (POD) and the interval within which a single laboratory's POD may lie.

The statistical treatment of results needed to provide estimates of method performance, principally the POD for a given concentration of target compound(s) and how much this probability might vary, differs from that used for quantitative methods. One example is the relation between the average POD across testing sites. The reproducibility standard deviation of the POD and the interval within which we can expect testing site PODs to lie produced by collaborative trials are not the same as the analogous relations for measurement results produced by quantitative methods.

Three recent publications give guidance on the statistical treatment of results for the validation of qualitative methods of analysis which are based on examining the POD: Probability of Detection (POD) as a Statistical Model for the Validation of Qualitative Methods (1), Probability of Identification (POI): A Statistical Model for the Validation of Qualitative Botanical Identification Methods (2), and How to Validate Qualitative Methods of Detection (3).

These publications give methods for analyzing the results of method validation using two important components of method performance: the average POD (1, 2) and the observed variation across testing sites (3). Approaches described in refs 1 and 2 are relevant when calculating the confidence limits of the average detection probability of the method in a single laboratory or a multi-site study. The approach described in ref. 3 estimates the confidence in the method's ability to reproducibly detect the analyte, given the observed variation in detection probabilities across all testing sites in the validation study. Method performance can be assessed using one or more of these approaches, depending on the type of information that is needed about the average POD (1, 2), or the interval within which PODs may lie (3; *see* Figure A1). If the between-testing sites variation in POD is large compared to the variation that comes from estimating the POD from a finite number of replicates, then the two approaches give complementary information about method performance. Where between-laboratory variation is small, each approach provides the same information.

Where a user needs to test their in-house method and they undertake an in-house validation, then the average POD (1, 2) observed during the validation study may be the most important measure of analytical performance for them. If they use the method to offer the analysis to customers, then assurance is needed that the analytical method meets a target for a POD on each instance of use. This can be achieved by an analysis of the validation study to estimate the interval within which the POD may lie on different days, using the statistical techniques described in ref. 3. In this example both approaches are needed to satisfy different stakeholders. Similar considerations apply to validation by collaborative trial or by using observed long-term performance.

The statistical methods described in refs 1–3 are designed to be accessible to as many users as possible. Where sufficient statistical expertise is available, a more accurate assessment of method performance may sometimes be achieved by fitting a model for the POD across analyte concentrations.

References

(1) 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 Int.* **94(1)**, 335–347

(2) LaBudde, R.A., & Harnly, J.M. (2012) *Probability of Identification (POI): A Statistical Model for the Validation of Qualitative Botanical Identification Methods*, *J. AOAC Int.* **95(1)**, 1–13

(3) Macarthur, R., & von Holst, C. (2012) *A Protocol for the Validation of Qualitative Methods of Detection*, *Anal. Methods* **4**, 2744–2754

ANNEX B
Considerations for the Design of Validation Experiments

The principal questions to be decided are number of concentration levels, matrices, number of test sites, and number of replicates per testing site. In general, the more levels that are studied, and the more replicates per level, the better the characterization of the POD curve. But there are tradeoffs that have to be made in order to strike a balance between confidence in results and having a practical, manageable study. Design of a practical multi-site experiment will require an understanding of the consequences of these tradeoffs with respect to the confidence intervals of the final results.

Number of test sites.—Historically, 10 valid data sets from collaborating testing sites have been required in order to validate a binary qualitative method in a multi-site collaborative study. The purpose of getting a large number of testing sites involved in the study is to get a wider subset of potential method users to contribute data to the study. Ideally, the chosen testing sites should be a random sampling of all potential method users. A larger subsample of testing sites will reduce the subsampling error and will mean the estimates that are obtained in the study will be less biased. In addition, with more testing sites, it will be easier to detect a testing site effect in the data, if one is significant.

Number of concentration levels.—Ideally, the experiment should verify that the method is sensitive to concentration in a general way, that at some low level, there is a low POD, and that at high concentration, the POD is high. The experiment will need to be designed to best characterize the POD curve, in as efficient a manner as possible. Generally, the minimum number of concentration levels to study should be three. There should be a very low concentration where the expected POD is close to zero, and if it is possible to obtain a sample with no analyte, then even better. This will demonstrate the method will not give a positive response at low, near-zero concentrations.

Second, there should be a high concentration, where the method is expected to give a very high percentage of positive responses. This will demonstrate that there is a concentration where the method responds to the target compound(s). Finally, there will be some concentration level where the POD is expected to be in a marginal range (0.25–0.75), which is important to identify so that the response curve can be better characterized and the transition concentration from low POD to high POD can be identified.

Alterations to the above basic scheme may be advised. More levels could be added in the marginal range to increase the confidence in estimation of the detection limit of the method. In some special cases, only two levels are studied. For example, if the high concentration is deemed to be more important, many replicates at the high concentration may be performed at the expense of replicates at the low level in order to focus the confidence interval of the high concentration estimates. *See AOAC INTERNATIONAL Methods Committee Guidelines for Validation of Biological Threat Agent Methods and/or Procedures* (OMA *Appendix I*) for examples of these types of designs.

It should be noted, if the purpose of the experiment is to compare the responses of two or more methods, then the emphasis should be placed in the marginal POD region, as this is the area where it will be most likely to discover differences in method responses. Refer to *AOAC INTERNATIONAL Methods Committee Guidelines for Validation of Microbiological Methods for Food and Environmental Surfaces* (OMA *Appendix J*) or ISO 16140 for examples of these types of validation experiments.

Number of replicates.—The number of replicates per level per testing site will determine the size of the confidence intervals of the POD estimate. The more replicates, the tighter the confidence interval will be. Also, simulation studies have shown that more replicates will lead to confidence intervals that are more accurate. In general 12 replicates per testing site is ideal, but eight replicates per testing site will be adequate given 10 testing sites are participating in the study.

Blind test portions.—It is very important that the samples and test portions provided to participants be blinded, so the collaborators cannot determine the expected outcome of any individual analysis. For example, if a study has three levels and eight replicates per testing site per level, the testing site would need to receive 24 test vials and be asked to analyze each vial independently, and the vials should be randomly coded so that the operator cannot distinguish the sample replication scheme.

Also, it is a good idea to randomly mix all levels together in the sample set so that it is equally likely to get a positive or negative response. The issue with studies at a single high or low level is that if all the results are expected to be positive or negative, it can be difficult to get good, unbiased results. So any study that focuses on a high level, should have at least 20% blank test portion added to the set as a check. It is important for the collaborators to understand that the experiment is designed with both positive and negative samples and the study is intended to test the method's ability to discriminate the two, and that any random sample in the set could give either a positive or negative response.

Statistical considerations for specified POD at a single concentration.—Many times when considering replicates per level, a desired POD is expected at a certain concentration level. For example, it may be hoped that the average POD may be expected to be at least 0.95 at the highest level studied. In some cases, the experiment may be designed so that the lower bound on the confidence interval of the POD estimate will be greater than 0.95 if a certain number of replicates are analyzed and a certain specified number of the replicates are positive. For this example, to assure at least 95% confidence that the true POD level is above 0.95, run 96 replicates total and have at least 95 positive results. This allows for one negative in the set without striking the whole study. For the same confidence with no allowed negative results, run 60 replicates. Other schemes can be devised given a maximum/minimum POD value and a confidence intervals and allowing various numbers of nonconforming results. *See* LaBudde, R.A., & Harnly, J.M. (2012) *Probability of Identification: A Statistical Model for the Validation of Qualitative Botanical Identification Methods, J. AOAC Int*. **95**, 273–285.