AOAC INTERNATIONAL
Presidential Task Force on
Best Practices for Microbiological Methodology

US FDA Contract #223-01-2464, Modification #12
Task Force Report

I. Background

During the past several years, issues have been raised about the limitations of the current AOAC guidelines for validation of microbiological methods. These issues have included the high rate of apparent false negative results when unpaired samples are used, the lack of a definitive acceptable range for "fractional positive" results for qualitative studies and the lack of appropriateness of the guidelines to bacterial toxins. A statistical task force was formed in 2003 to try to address the statistical issues, especially in the case of unpaired samples, and propose solutions. A set of recommendations was drafted, but as yet the recommendations have not been adopted by the Official Methods Board of AOAC. This task force did not address all the issues and concerns previously raised relative to validation of microbiological methods, but focused on the issues of importance to the US FDA as outlined in the objectives of the contract.

Modification #12 of USFDA Contract #223-01-2464 arose from discussions of the limitations of the current AOAC microbiology guidelines and a proposal to re-evaluate the AOAC guidelines was created. Modification #12 of the contract is focused on developing recommendations on the best practices for validation of microbiological methods by an international team of experts. The goal of the group was to consider the technical and statistical aspects of the current AOAC guidelines and ISO 16140 and to recommend new approaches where needed, without regard to harmonization, consensus within AOAC INTERNATIONAL or consensus among international validating organizations.

To lead the project, AOAC appointed Russ Flowers to Chair the Presidential Task Force on Best Practices for Microbiological Methods (BPMM, hereafter referred to as Task Force). A task force structure quickly took shape, comprising a Steering Committee (SC) of key individuals with varying expertise and four Working Groups – Detection Limits (DLWG), Matrix Extension (MEWG), Sampling (SAWG), and Statistics (STWG). The working groups were chaired by Steering Committee members and populated by international experts in that topic area. Great effort was expended to identify technical experts from government, industry, reference laboratories, and academia with varied backgrounds in food safety, quality assurance, clinical diagnostics, veterinary diagnostics and engineering. Not surprisingly, some of these experts also serve on committees for other standards organizations, such as ISO (International Standards Organization), ASTM (American Society for Testing and Materials), CLSI (Clinical and Laboratory Standards Institute, formerly NCCLS), and CEN, the European Committee for Normalization. Care was taken, however, to select scientists and technical experts, without introducing political agendas.

It is interesting to note that ISO TC 34 SC 9 is also considering a revision of the ISO 16140 guidelines and the recommendations from the BPMM task force will provide valuable input to that process. The BPMM project is an important step in the international harmonization of microbiological methods.

The objectives contained in Modification 12 of the US FDA contract were assigned to the working groups as appropriate, and the task of the working groups was to address the objectives by developing recommendations based on sound scientific and statistical principles. The Steering Committee provided guidance to the working groups and served as editors for the final reports. There were some topic areas that overlapped between Working Groups and, therefore, for future publication purposes, the ideas and recommendations of the task force would be best organized by topic area rather than by contract objectives as contained in this report.

This report summarizes the recommendations of the task force and is supported by appended working group reports, which provide the details behind the recommendations. The goal of the BPMM task force, in the short period of time allotted for the contract, was to determine the best practices for validation of microbiological methods and to make recommendations for consideration and further research by AOAC and US FDA. The Task Force did not attempt to create new microbiology validation guidelines as many of the recommendations represent new approaches that must be further evaluated from the perspective of practical application. There is no expectation of adoption of the recommendations by AOAC INTERNATIONAL. After discussing the merits and limitations of the BPMM recommendations, it is hoped that additional work will be funded to further refine and practically evaluate the recommendations presented herein. The Steering Committee recommends first that existing data be used to compare the statistical recommendations to current practice, and then laboratory feasibility studies be conducted to test proposed study designs and sample preparation techniques. These additional efforts would be expected to lead to development of new detailed guidelines for validation of microbiological methods that will be proposed for adoption by AOAC.

II. Executive summary

The Presidential Task Force for Best Practices in Microbiological Methods (BPMM) makes the following recommendations relative to the objectives of Contract #223-01-2464, Modification #12. A more complete explanation and justification for the recommendations is given in the appended documents. A glossary of terms is found in Appendix O.

Objective 1: Once a microbiological method has been validated for an array of specific foods and specific strains of a microorganism:

a) *To what extent can these results be extrapolated to other foods and other strains?*
b) *Are there abbreviated but scientifically/statistically appropriate procedures/protocols by which a validation can be expanded to include additional foods and/or strains?*
c) *How can methods be applied to specific foods, where no validation has been performed?*

The BPMM Task Force recommends new food sub-categorization schemes based on proximate analysis, level and types of background microflora, presence of inhibitors and other characteristics of food matrices that may affect microbial growth, recovery or analytical procedures. Based on the new scheme, varying degrees of verification or validation (from no verification or validation to harmonized collaborative validation) would be required in order to apply a method to a new food matrix. The degree of validation or verification is dependent on how closely related the new matrix is to previously validated matrices and on the current validation status of the method (single lab validated, multiple lab validated or harmonized collaboratively validated).

A list of essential reference organisms and toxins was compiled to address the issue of variability of strains. The organisms and toxins represent antigenic and genomic variability and are recommended to be used as part of the inclusivity testing as appropriate for the method target. Other food-borne isolates should be added to the inclusivity list based on the claimed application of the method.

Objective 2: What are the scientific/statistical bases for developing performance standards against which the validation of methods should be based?

The Task Force recommends that performance standards be based on public health objectives (PHOs) and/or fitness for purpose criteria. In general, statistical methods should be used to assist in setting realistic performance standards. These procedures should be based on control of error related to a true negative testing positive (Type I) and error related to a true positive testing negative (Type II). Levels of poor performance that must be detected (with stated probabilities) should be determined. Appropriately determined sample sizes should be used to meet the stated goals. This approach would be a change from current practices where studies are accepted on the basis of standard designs for number of laboratories, materials, and replicates, and standard criteria for suitability of the summary statistics. The design specifications and resulting reliability

estimates should form the basis of applicability statements for test and measurement methods.

Objective 3: What are reasonable performance standards [criteria] when microbiological methods are to be validated for use for: 1) Attribute (presence/absence) testing (for both 2-class and 3-class sampling plans), 2) Variables (quantitative) testing of batches, and 3) Process control testing of processes or cross-batch testing?

Whether the method is to be used for attribute or variables testing, performance standards are similar. Ruggedness tests should be performed on the analytical procedure being used. The validation of a test should include estimates of sensitivity, specificity and accuracy. Reproducibility and repeatability should be determined through a detailed collaborative study and ranges of these measures should be published for quality assurance purposes. Results reporting should include a 95 % confidence interval rather than a point estimate of the mean. More detailed and specific recommendations are given in Appendix C.

Recommendations relative to Statistical Process Control (SPC) include Shewhart Charts of control samples with statistical control limits. Standard rules for setting control limits and evaluating control of these charts with respect to Type I and II errors should be followed. Specification limits should not be part of a SPC system. Further details of the SPC recommendations are described in more detail in Appendix F.

Objective 4: What are the scientific/statistical bases for determining the lower limit of detection for microbiological methods? How is the lower limit of detection validated during the validation of a method? How is the relative performance of a method determined as the lower limit of detection is approached and what is the best way of characterizing this performance?

The detection limit for qualitative tests is best described as the "LOD_{50}", or number of organisms per gram of sample at which 50% of the tests are positive. "LOD_{50}" not used in the analytical chemistry sense of LOD and LOQ. It is used in the microbiological sense of an endpoint where the methods are able to estimate around the level of a few particles (bacterium, virus, or genetic macromolecule) per analytical portion. This is possible because virtually unlimited amplifiability of such particles is possible due to their ability to multiply themselves in appropriate conditions. Fifty percent endpoint calculation methods allow for failures to inoculate resulting from imperfect homogeneity at low particle numbers per analytical portion. Such calculations do not assume or require paired samples. LOD_{50} is determined with a nonparametric (distribution free) version of probit analysis, and an experimental study using at least 4 dilutions in which at least two of the dilutions have "fractional positives" in order to estimate better the $LOD_{50.}$ Estimates of other percentiles, such as the LOD_{90} (number of organisms per gram of sample where 90% of results are positive) may be possible in the future development of this approach. The LOD_{50} procedure requires that one dilution level has 0% positive results and one dilution level has nearly 100% positive results (allowing for measurement error in the test laboratories). The associated confidence limits refer to the uncertainty of

the estimated LOD_{50}. As proposed here, the LOD_{50} would be calculated for the pool of results from a multi-laboratory study with or without removing outlying laboratories. However, LOD_{50} values could be calculated for each participating laboratory for the purpose of removing outlying laboratories from a study.

The LOD_{50} approach assures that the lower limit of detection for a method is described. However, the number of organisms/g at the limit of detection must be determined on the day of analysis of the test portions. This can be accomplished by analyzing the seed (inoculated or naturally contaminated product diluted into the test samples) on the day of analysis, or by employing a reference method, if one exists, with a known limit of detection in that matrix. In a single method validation, where no reference method exists, the number of organisms/g at the limit of detection must be measured or calculated from measuring the level of organisms in the seed on the day of analysis. Methods for measurement are described.

For quantitative methods, the committee recommends use of the ISO 16140 procedure, which presents limits of detection and quantification as functions of the variability of blank (or very low) samples. The committee recognizes, however, that alternative procedures exist that should be investigated, such as the ISO 11843 series on capability of detection, or the nonparametric analog of that procedure described in the CLSI document EP17-A on Limits of Detection and Quantitation. These procedures recognize the importance of Type I and Type II errors, and that variances of signals from truly negative and truly positive samples can be different. There are related strategies for designing experiments to use the ISO/CLSI approach.

Objective 5: What are the scientific/statistical bases for developing validation protocols that adequately take into account the biological variation that exists within both the microorganisms and toxins produced by these microorganisms for which methods are developed and the foods which will be analyzed?

See Objective 1 for discussion of biological variation of microorganisms and categorization of foods.

Validation of methods for toxins produced by microorganisms present a different set of challenges than validation of methods for microorganisms themselves. It is strongly recommended that methods targeting toxins be validated according to the AOAC chemistry guidelines and reviewed by the Chemistry and Microbiology Methods Committees. The methods for preparation of the microbial toxins, dilution into the food matrix and end-user sampling plans may need to be defined by consulting microbiologists, but the validation protocol appropriate for other chemical contaminants should apply.

Objective 6: What are effective means for articulating the uncertainty associated with microbiological methods?

Uncertainty in measurements using quantitative procedures is best estimated following an all-inclusive, or "top down" approach. This approach does not attempt to estimate all components of uncertainty separately and it does not require a detailed mathematical model of how those components are combined (ISO 19036). This approach is in contrast to a "bottom up" approach, which requires estimation and combination of variances at all stages of an analysis. This cannot be done routinely, however, so standard, or assumed, variances are used which align the combined estimate to the basic method rather than the analytical result. The "bottom up" approach is likely to underestimate uncertainty due to sources of uncertainty that are not considered. By contrast, the "top down" approach makes no attempt to set generic estimates of uncertainty for specific test methods and rightly aligns the estimate of uncertainty with a specific analysis (or set of analyses). The "top down" approach is consistent with the Guide to the Expression of Uncertainty in Measurement (GUM (2000), Quantifying uncertainty in measurement, BIPM/IEC/IFCC/ISO/IUPAC/OIML, published ISO) principles that allow combination of sources of uncertainty that are difficult to estimate individually.

For qualitative methods, measurement uncertainty for the result cannot be expressed directly – instead, the observed effect is on the probability of reporting an incorrect result. This can be estimated with false negative and false positive rates, for those methods with confirmation procedures. For some measurement procedures, uncertainty can be expressed as the standard error of the LOD_{50}, as estimated by the Spearman-Kärber method. This procedure estimates uncertainty where it is most important, which is at the border of the determination of "present" or "absent". The work of ISO Technical Committee 34, Subcommittee 9 is not yet completed, so the STWG recommends active participation in the efforts of this subcommittee.

Objective 7: How is the statistical basis of a method influenced if the homogeneity of the sample cannot be assumed, particularly at the very low CFU level? How does this influence the performance parameter of a method? How can samples be prepared to minimize this effect? Define optimum procedures for sampling.

With regard to the statistical validity of low level contamination, a supplemental statistical treatment is presented. This technique, LOD_{50}, is not suggested as a replacement for existing tests for significant differences, e.g. Chi Squared, but rather offered as a data treatment that could provide some measurement of the potential variability associated with low level contamination. This may be particularly relevant given that the LOD Working Group believes that low level contamination of matrices, whereby fractional recovery of positives samples occurs within an inoculation level, is the preferred method for defining assay performance.

Even though the homogeneity of the sample cannot be assumed, protocols are presented to minimize this impact. Furthermore, specific protocols are recommended for different categories of food matrices (high moisture food and low moisture food).

Objective 8: Can a 2-dimensional classification matrix be developed using (1) rating of importance/urgency of intended use and (2) degree of validation, as the dimensional

factors? Examples of intended use include: (a) response to a recently emerged microorganism, (b) process control, (c) regulatory screening, (d) regulatory confirmation, and (e) forensic attribution. Examples of validation include: (a) published paper, (b) single laboratory validation, (c) multiple laboratory validation, and (d) full collaborative validation.

Using the fit for purpose concept, the task force recommends varying degrees of validation for microbiological methods, depending on their intended use (see Appendix M). The study design (number of levels and number of replicates) and type of validation [single lab (SLV), multi-lab (MLV) and harmonized collaborative (HCV), or variants on these per specific design recommendations] will be dependent on whether the method is intended for widespread use, such as a screening method, or regulatory use in one or a few labs and what level of statistical confidence is required by the end user for that intended use. The degree of uncertainty that can be tolerated will depend on many factors; including urgency, cost, availability of confirmatory methods, laboratory (or field) analytical capabilities, etc. However, the degree of confidence required for regulatory, legal and forensic applications will certainly require the highest level of analytical confidence in the data, but may not be as demanding in terms of speed, ruggedness, reproducibility (inter-laboratory variance) and cost per test. Compiling a detailed set of recommendations requires further research and input from end users to define the limits of acceptable performance for each application.

Objective 9: What are the minimum performance criteria (percentage correct answer on known controls with defined confidence limits) for each factor listed in Objective 8?

Ideally, performance specifications or acceptance criteria should be based on risk analysis and historical analytical capabilities. In practice, however, developing statistically derived (comparative where available) performance characteristics through validation studies (SLV, MLV, HCV, or studies designed for the intended purposes and needs of the method as recommended by the BPMM study) is more reasonable, allowing potential users of the method to determine its application based on the fitness for purpose concept.

As the number of laboratories increases, the apparent dispersion in results will increase, but properly constructed confidence limits for performance measures will decrease due to having better estimates of the largest source of variability, the inter-laboratory variability. Clearly written package inserts, detailed validation protocols and method training are key factors to controlling this variability in the collaborative study and in the end-user application of the method.

The task force recommends that the level of confidence for different applications of methods be defined using the fit for purpose approach (intended use and end user requirements), and then appropriate validation study designs and verification criteria can be developed within the constraints of practicality. For example, regulatory agencies may determine that a method for detection of *Listeria monocytogenes* in food must have an LOD_{50} of 1-3 CFU/25g at 95% confidence in the single laboratory study. The SLV study design and acceptance criteria would be based on this target value and variance.

Likewise, acceptable inter-laboratory variances can be used to design multi-lab and collaborative studies and set acceptance criteria for these studies. In reality, the needs of the end user and the practicality of the study design must be balanced.

A retrospective analysis of current AOAC, OMA and PTM methods using the LOD_{50} approach would provide a starting point for determining target performance criteria for various intended uses. A corollary to this recommendation is that OMA precollaborative studies must be published as these studies generally provide SLV performance characteristics for a wider variety of matrices than collaborative studies.

Objective 10: What are the appropriate statistical tools to be used for interpretation of validation studies?

The Statistics Working Group recommends the use of robust statistical procedures that are not as severely affected by extremely large or small results that can be misleading with more conventional procedures. It also recommends against the removal of outliers from collaborative studies, except for assignable causes. The group recommends review of instances of laboratories in a collaborative study that give indications of having a different application of a method, to see if the method is clearly defined. The committee prefers strong cautions about the concept of "false negative" and "false positive" results due to the vagaries of microbial distribution, the difficulty of confirming all positives and negatives, and the likelihood of misinterpretation. Alternative confirmation procedures should be considered, such as nucleic acid testing. Any estimates of "sensitivity" for low level samples should be corrected statistically for the assumed number of true negatives, based on an assumed Poisson distribution of organisms in the samples.

Chi Squared analysis according to McNemar for paired samples is recommended where a reference method is available. An alternative formula for Chi Squared analysis of unpaired samples should be considered.

It is important that collaborative (interlaboratory) studies be analyzed carefully. The group recommends that current practices of deleting statistical outliers be replaced with a procedure to investigate laboratories that perform differently for an analyte, to see whether the cause can be explained, often because the laboratory had an incomplete understanding of the method. In these cases the method needs to be described better.

Objective 11: What are the test variables (e.g., number of strains, foods, inoculum levels) that should be considered for each of the factors listed in Objective 8?

The basic elements of validation studies include inclusivity and exclusivity, characterization of the method performance, and, where applicable, estimate inter-laboratory variation. Additional elements include ruggedness, stability, lot-to-lot variation, and instrument variation (if applicable). The test variables to consider in the design of a validation study include:

- Intended use
- Confidence required for intended use

- Number of inoculum levels
- Number of replicates per level
- Number of labs
- Food claim (from single matrix to multiple categories)
- Analyte claim (Genus, species, or strain)

Objective 12: Can acceptance criteria be established for methods modification/substitution?

It is logical to say that when a method is modified, its performance should be at least as thoroughly evaluated as was the original method. However, recognizing that the modification of a method may have benefits other than enhanced performance parameters, a modified method cannot be required to perform better than the original. Further, since there are many applications for methods (screening, regulatory action, process control, etc.) a modified method used for a different application may be acceptable even though some of its performance characteristics may be inferior to the original method. For example, increased sensitivity or broader inclusivity for a "screening method" may result in poorer specificity and/or exclusivity compared with the original method. Therefore, the acceptance criteria for method modification must be based on the claim being made (broader inclusivity, enhanced exclusivity, increased sensitivity, faster time to result) and the intended use of the method (screening, confirmation, process control, etc.).

Objective 13: Define performance criteria for discrete vs. attribute testing methods.

This objective was removed from the contract on July 21, 2005, following a request for clarification on April 18, 2005. The BPMM and contractor agreed that this objective is sufficiently covered under Objective 3.

Additional Recommendation

As briefly mentioned above in response to objective 3, the task force recommends that ruggedness testing be included as part of every microbiological method validation, similar to what is currently done in the AOAC Research Institute *Performance Tested Methods*^{SM} program. Ruggedness testing involves the deliberate introduction of minor variations in a method procedure. These minor variations should be of a magnitude that might be expected to occur in the hands of the end user. Parameters to be tested might include reagent volumes, reaction temperature and time, enrichment temperature and time, and the like. The specific parameters to be varied would depend on the test method technology and type (quantitative or qualitative method; bacterial, viral or toxin method) and would be determined on a case-by-case basis. Ruggedness testing would target those parameters deemed most critical to method performance in order to provide guidance to the end user regarding the control of those parameters. Ruggedness testing is included in the plans for future research.

Future Research

Many of the recommendations and ideas of the task force require further review and development. For a description of the suggested areas for future research, see Appendix N.

III. Appendices

A. Detection Limits WG Report
B. Matrix Extension WG Report
    1. Matrix Extension Essential Organisms List
C. Sampling WG Executive Summary
D. Sampling WG Introduction
E. Sampling WG Enclosure A – Measurement Error
F. Sampling WG Enclosure B – Statistical Process Control
    1. Appendices for Statistical Process Control
G. Statistics WG Executive Summary
H. Statistics WG Report Part 1 – Developing Standards and Validating Performance
I. Statistics WG Report Part 2 – Study Variables
J. Statistics WG Report Part 3 – Uncertainty
K. Statistics WG Report Part 4a – LOD_{50}
L. Statistics WG Report Part 4b – LOD_{50} Spearman-Kärber Worksheet
M. Classification Matrix
N. Recommendations for Future Research
O. Glossary
P. Task Force Membership
    1. Steering Committee
    2. Detection Limits Working Group
    3. Matrix Extension Working Group
    4. Sampling Working Group
    5. Statistics Working Group
    6. AOAC Staff