Developing standards and validating performance: scientific/statistical bases for describing the validation of performance.

I. Principles:

1. Performance standards should be based on appropriate statistics for describing method performance and expert/regulatory objectives for the intended use. Performance statistics might vary for different measurement technologies, although attempts should be made to harmonize these. Minimal performance criteria could be different for every performance statistic and for every use.

2. The performance characteristics should be estimated with experimental protocols to assure that confidence intervals for the statistics will be small enough for their intended use (this controls Type I error and Type II error).

3. All performance characteristics should be specific for a defined organism or strain of particular interest. That is, the "measurand" (or "analyte" in common usage) should be defined exactly; sometimes this might include a single strain, sometimes a class of strains, a genus, or a group of organisms. Similarly, the measurand might be a specific microbial toxin, a group of toxins, or some other parameter. With a carefully described measurand, the concepts of "inclusivity" and "exclusivity" are variants of "sensitivity" and "specificity". These statistics are useful for general descriptions of a method that is approved for a microorganism with many important strains.

4. Results from experiments in a single laboratory can be useful for the design of collaborative studies, but should not be used alone to establish claims for method performance (except for the particular laboratory's own purposes). Performance characteristics should be estimated, wherever possible, with experiments conducted in two or more laboratories that have demonstrated competence with this type of microbiological procedure and experimental protocol. At least one of the laboratories must be independent of the manufacturer/developer of the method. It is acceptable to validate performance of a method in a single laboratory and have a second laboratory verify the performance on a carefully selected subset of matrices, but this needs to be done with care, and following the recommendations from the BPMM Matrix Extension Recommendations or ISO 16140. Therefore determinations of bias between an alternative and reference method cannot be determined in a single laboratory, nor can inclusivity or exclusivity (sensitivity and specificity), unless the determinations are verified in at least one other laboratory.

5. It is essential to differentiate between the uncertainty (*sic* lack of precision) of a method and the uncertainty of an estimated value derived using that method (*sic* the measurement uncertainty). Estimates of measurement uncertainty should be derived from appropriate "top down" procedures using intra- or inter-laboratory randomized trials; in some instances, such estimates may be specific to each individual laboratory undertaking a specific test. By contrast, uncertainty estimates for the method are derived by "bottom up" procedures and must be used with care since they normally underestimate the true extent

of uncertainty of a measurement. Upper limits for uncertainty estimates may be used by those laboratories that can demonstrate competence with the method.

II. Considerations for Statistics and Statistical Methods

1. The statistics used should possess the following qualities:
   - Unbiased, maximum likelihood estimates for the performance characteristics of interest.
   - Appropriate for the distribution of data from which they will be calculated.
   - Understandable and intuitive for microbiologists and regulators.
   - Sensitive to the most common sources of error or deviation from expected performance.

2. Criteria for the suitability of the performance statistics should be based on the following considerations:
   - Professional judgment on the performance level required for the method for the intended use. This should be based on considerations for public health or fitness for purpose, and technical knowledge of the method.
   - There should be definitions of performance that is not suitable for a prescribed purpose; that is, poor performance that should be detected with high probability.
   - Probability of improperly rejecting a method as unsuitable, when in fact it is suitable for the intended use (control of Type I error).
   - Probability of improperly accepting a method as suitable, when in fact it is not suitable for the intended use (control for Type II error).

3. The data used to characterize performance should be based on the following:
   - Data from more than one laboratory.
   - Statistics based on all results received from competent laboratories, all following the same well-defined instructions for the measurement procedure and reports (discard only those data outliers for which there is a known cause).
   - Data transformed to reasonable normality and analyzed using appropriate robust or nonparametric methods. Severe non-normality of the transformed data (many statistical outliers) or evidence of bimodality should be resolved prior to analysis of the data.

III. Considerations for calculating performance statistics from collaborative studies.

Carefully designed collaborative studies are preferred for describing the performance capabilities of a measurement procedure. The guidelines for determining the numbers of laboratories, levels, and replicates are well established (see for instance McClure & Lee, 2005). Procedures for analysis of the data are less well established.

1. Before summary statistics are generated, it is important to look first for laboratories that seemed to have difficulties with more than one sample, or whose results are consistently high, low, or highly variable across levels. These are the laboratories that were possibly affected by ambiguous instructions, a missing step in the procedure, or other inherent weakness in the measurement procedure. These situations must be investigated before data

analysis proceeds. Any truly erroneous results must be eliminated - or in some cases they can be corrected (as in decimal point errors or switched samples). No results should ever be eliminated for purely statistical reasons. If reasons cannot be found, then the variability is assumed to be representative of the procedure. Obvious any bimodality in the data must also be resolved, possibly using 'bump hunting' procedures. Once the truly erroneous results are eliminated then the statistical processing can commence.

2. ISO 16140 recommends use of robust statistical procedures rather than conventional parametric statistical techniques, and the BPMM STWG agrees with this recommendation. However, whether extreme results are eliminated as outliers or have their impact limited with robust techniques is less important than the analyst's investigation of how such outlying results occurred.

3. For qualitative method comparison studies, the definition of the "Reference Method" is important for naming the performance statistic. If the Reference Method is definitive for confirming the presence and absence of an organism, then it is possible to use "false positive" and "false negative" as summary measures for an alternative method. Similarly, if definitive confirmation techniques are available then "false positive" and "false negative" for an alternative method may be reported. However if the Reference Method is not definitive, then performance measures are relative to the Reference Method and must be described as "relative sensitivity" and "relative specificity".

4. McNemar's Chi-Square test is appropriate for testing for significant disagreement between Reference and Alternative Methods, but is appropriate only when samples are truly pairs – that is when they share a common enrichment or pre-enrichment step. Artificially linked samples are not appropriate for McNemar's test.

5. The term "false negative" (FN) is a confusing concept, even when using only confirmed positives. When there are few organisms in the sample, the FN rate may be a combination of results where an organism was present but not detected and results from samples that truly contained none of the target organisms, due to inhomogeneous distribution of organisms in the larger sample. It is possible using the Poisson distribution (or if appropriate the Binomial or Negative Binomial distribution), to adjust the false negative rate to account for the estimated number of true negatives. Therefore, even when only confirmed positives are used, false negative rates should be adjusted for the theoretical likelihood of having a true negative.

6. The LOD_{50} is an independent descriptor of performance, and is preferred to measures that are relative to the Reference method (such as false negative or false positive). .

7. If possible, all samples should have their positive/negative status confirmed by an independent methodology (except for those samples that are positive by both the reference and alternative methods). Samples that are negative by both methods should also be confirmed by independent methodology, if possible.

IV. Recommended performance statistics

Qualitative Methods

1. Number of cfu per gram of matrix for 50% probability of a positive signal (LOD_{50})
2. Number of cfu per gram of matrix for 90% probability of a positive signal (LOD_{90})
3. Probability of a negative signal when the Reference Method indicates no organisms is present (relative specificity).
4. Probability of a negative signal when common contaminants, but not the target organism, are added to a sterile sample (specificity).
5. Probability of a positive signal when the Reference Method indicates organisms are present (relative sensitivity).
6. Proportion of replicates with same result (repeatability)
7. Proportion of results from different laboratories with the same correct result (reproducibility).
8. Standard Error of the LOD_{50}, for use in estimating the effect of measurement uncertainty on the probability of obtaining an incorrect result.

Quantitative Methods

1. Difference between replicate samples obtained under repeatability conditions (intra-laboratory repeatability).
2. Difference between replicates from the same material in the same laboratory, using changed conditions (intermediate reproducibility).
3. Difference between average results from different laboratories, testing the same material (reproducibility).
4. Average difference between the Alternative Method and the Reference Method pooled across multiple competent laboratories (relative method bias).
5. The extent to which the measurement signal is proportional to the number of organisms in the sample (linearity).
6. Range of quantification: the lowest and highest signals that can be detected with adequate uncertainty, obtained by dilution. For plate count methods, this is the range of counts per plate where results can be obtained with a stated degree of repeatability precision.
7. Lowest level where results can be obtained with a stated uncertainty that is fit for its purpose (limit of quantitation).