TECHNICAL REPORT

ISO/TR 13843

First edition
2000-06-01

Water quality — Guidance on validation of microbiological methods

Qualité de l'eau — Lignes directrices pour la validation des méthodes microbiologiques

Reference number
ISO/TR 13843:2000(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 734 10 79
E-mail copyright@iso.ch
Web www.iso.ch

Printed in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this Technical Report may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 13843 was prepared by Technical Committee ISO/TC 147, *Water quality*, Subcommittee SC 4, *Microbiological methods*.

Water quality — Guidance on validation of microbiological methods

1 Scope

This Technical Report deals with validation of microbiological methods, with particular emphasis on selective quantitative methods in which the quantitative estimate is based on counting of particles either directly, with the aid of a microscope, or indirectly, on the basis of growth (multiplication) into colonies or turbidity.

The principles and procedures within this scope are commonly known as the presence/absence (P/A), most probable number (MPN), colony count and direct (microscopic) count.

This Technical Report does not apply to the validation of the so-called rapid or modern methods which mostly depend on measuring products or changes due to microbial activity but do not address the detection of individual particles.

2 Terms and definitions

For the purposes of this Technical Report, the following terms and definitions apply.

2.1
accuracy of measurement
closeness of the agreement between a test result and the accepted reference value

NOTE The term “accuracy”, when applied to a set of test results, involves a combination of random components and a common systematic error or bias component.

[ISO 3534-1:1993, 3.11]

2.2
analyte
measurand
particular quantity subjected to measurement

NOTE 1 See reference [5].

NOTE 2 In microbiology the analyte is ideally defined as a list of taxonomically defined species. In many cases, in practice the analyte can only be defined by group designations less accurate than taxonomic definitions.

2.3
analytical portion
test portion
volume of particle suspension inoculated into a detector unit

NOTE Examples of a detector unit are agar plate, membrane filter, test tube, microscopic grid square.

2.4
application range
range of particle concentrations routinely subjected to measurement by a method

2.5
categorical characteristic
method performance characteristic numerically expressed as a relative frequency based on P/A or +/– classification

2.6
CFU, deprecated
colony-forming unit, deprecated
CFP, deprecated
colony-forming particle, deprecated

NOTE The term was originally introduced to convey the idea that a colony may originate not only from a single cell but from a solid chain or aggregate of cells, a cluster of spores, a piece of mycelium, etc. It mistakenly equates the number of colonies observed to the number of living entities seeded on the medium. Growth unit, viable particle, **propagule** (2.27) and **germ** (2.13) are terms with similar meanings but convey the original idea better and apply not only to colony-count methods but also to MPN and P/A.

2.7
coefficient of variation
CV
relative standard deviation
for a non-negative characteristic, the ratio of the standard deviation to the average

NOTE 1 The ratio may be expressed as a percentage.

NOTE 2 The term "relative standard deviation" is sometimes used as an alternative to "coefficient of variation", but this use is not recommended.

[ISO 3534-1:1993, 2.35]

NOTE 3 In this Technical Report the term coefficient of variation (CV) is used when the relative standard deviation is expressed in percent (CV % = 100 RSD).

2.8
collaborative test
method or laboratory performance test where several laboratories join in an experiment planned and co-ordinated by a leader laboratory

NOTE Collaborative tests are mainly of two types. Intercalibration exercises are made to allow laboratories to compare their analytical results with those of other participating laboratories.

Method performance tests produce precision estimates (repeatability, reproducibility) out of data accumulated when several participating laboratories study identical samples with a strictly standardized method.

2.9
confirmed [verified] colony count
x
presumptive colony count corrected for false positives

$$x = pc = \frac{k}{n}c$$

where

c is the presumptive count;

p is the true positive rate;

n is the number of presumptive positives isolated for confirmation;

k is the number confirmed.

2.10
control chart
two-dimensional scattergram for monitoring method performance with control values obtained by a Type A study

NOTE In control charts the horizontal axis is usually in the time scale or ordinate scale and the control variable is the mean or some precision measure (s, CV, RSD).

2.11
detector
particle detector
plate of solid matrix or a tube of liquid containing a nutrient medium for counting or detecting living microbial particles

2.12
detection set
detector set
combination of plates or tubes on which quantitative estimation of microbial concentration in a sample is based

NOTE The detection set is the set of plates or tubes utilized for numerical estimation of a single value.

EXAMPLES Parallel plates of a suspension, plates from consecutive dilutions, 3 × 5 tube MPN system, microtitre plate.

2.13
germ
living entity capable of producing growth in a nutrient medium

cf. **propagule** (2.27)

2.14
guidance chart
two-dimensional scattergram for presenting method performance data (quantity or precision) with arbitrary guide values or guide values obtained by Type B reasoning

NOTE In guidance charts, the horizontal axis is usually the colony count per detector.

2.15
heterogeneous Poisson distribution
distribution arising when the mean of a Poisson distribution varies randomly from occasion to occasion

NOTE 1 See reference [11].

NOTE 2 See also negative binomial distribution (2.19).

2.16
limit of detection
particle number x (per analytical portion) where the probability p_0 of a negative result equals 5 %

NOTE 1 Probability of a positive result $p(+) = 1 - p_0$.

NOTE 2 a) Calculation of x via Poisson distribution:

$$x = \ln\left(\frac{1}{p_0}\right) = \ln\left(\frac{1}{0,05}\right) = \ln(20) = 3,00$$

b) Calculation of x via negative binomial distribution:

$$x = \frac{\left(p_0{}^{-u^2} - 1\right)}{u^2} = \frac{0,05^{-u^2} - 1}{u^2} = \frac{20^{-u^2} - 1}{u^2}$$

2.17
limit of determination
lowest average particle concentration x per analytical portion where the expected relative standard uncertainty, equals a specified value (RSD)

NOTE a) Calculation of x via Poisson distribution:

$$x = \frac{1}{(\mathrm{RSD})^2}$$

b) Calculation of x via negative binomial distribution:

$$x = \frac{1}{(\mathrm{RSD})^2 - u^2}, \text{ given overdispersion factor} = u$$

2.18
linearity
linear dependence of the signal on concentration of the analyte

cf. **proportionality** (2.28)

2.19
negative binomial distribution
a particular "overdispersed" statistical distribution of counts

NOTE 1 Its variance can be expressed as

$$\sigma^2 = \mu + u^2 \mu^2$$

where μ is the mean.

NOTE 2 In this Technical Report the square of the overdispersion factor (u) is substituted for the inverse of the exponent ($1/k$) of the standard formula for the negative binomial distribution.

2.20
overdispersion
variation in excess of Poisson randomness

NOTE It is detected qualitatively by the Poisson index of dispersion, and measured quantitatively by estimating the parameter u (overdispersion factor) of the negative binomial distribution.

2.21
overdispersion factor
u
additional random uncertainty of determination in excess of the Poisson distribution, measured in terms of relative standard deviation

2.22
overlap error
crowding error
systematic depression of colony counts due to confluence of colonies

NOTE Quantitatively, overlap error depends primarily on the fraction of available growth space occupied by colonial growth.

2.23
parallel counts
particle or colony numbers in equal analytical portions drawn from the same suspension

2.24
Poisson distribution
fully random distribution of particle numbers when sampling a perfectly mixed suspension

NOTE The probability $P(k)$ of observing exactly k units in a test portion when the mean equals μ is calculated from

$$P(k) = \frac{\mu^k}{k!}e^{-\mu}$$

2.25
precision
closeness of agreement between independent test results obtained under stipulated conditions

NOTE Precision does not relate to the true value or the specified value. It is usually expressed in terms of imprecision and computed as a standard deviation of the test results.

2.26
primary validation
full validation
establishment of the specifications for the performance of a new method and/or experimental verification that a method meets theoretically derived quality criteria

2.27
propagule
a viable entity, vegetative cell, group of cells, spore, spore cluster, or a piece of fungal mycelium capable of growth in a nutrient medium

cf. **germ** (2.13)

2.28
proportionality
agreement of observed particle counts with the volume (or dilution) of a series of analytical portions from a common root suspension

NOTE Proportionality is computed for statistical evaluation as the log-likelihood ratio statistic G^2 with n–1 degrees of freedom.

2.29
qualitative method
method of analysis whose response is either the presence or absence of the analyte in a certain amount of sample

NOTE See reference [10].

2.30
recovery
general term for the number of particles estimated in a test portion or sample, with the understanding that there is a true (although unknown) number of particles of which 100 % or less are "recovered" by the detector

2.31
relative accuracy
degree of correspondence between the response obtained by the reference method and the response obtained by the alternative method on identical samples

NOTE See reference [10].

2.32
relative difference
d
difference between two measured values divided by their mean

$$d = \frac{x_A - x_B}{\bar{x}} = \frac{2(x_A - x_B)}{x_A + x_B}$$

$$d\,\% = 100\,d$$

NOTE For all practical purposes, the same value results from the calculation $d = \ln(x_A) - \ln(x_B)$.

2.33
relative recovery
ratio (A/B) of colony counts obtained by two methods tested on equal test portions of the same suspension, where B is the reference (when applicable)

2.34
relative standard deviation
RSD
estimate of the standard deviation of a population from a sample of n results divided by the mean of that sample

$$\text{RSD} = \frac{s}{\bar{x}}$$

cf. **coefficient of variation** (2.7)

2.35
repeatability
closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement

NOTE 1 See *Guide to the expression of uncertainty in measurement* [6].

NOTE 2 Repeatability is computed as $r = 2{,}8 s_r$, where s_r is the repeatability standard deviation.

2.36
reproducibility
closeness of the agreement between the results of measurements on the same measurand carried out under changed conditions of measurement

NOTE 1 See *Guide to the expression of uncertainty in measurement* [6].

NOTE 2 Reproducibility is computed as $R = 2{,}8\ s_R$,

where

s_R is the reproducibility standard deviation usually compounded from the between-laboratories standard deviation s_L and repeatability standard deviation s_r:

$$s_R = \sqrt{s_L^2 + s_r^2}$$

2.37
robustness
insensitivity of an analytical method to small changes in procedure

NOTE 1 See reference [23].

NOTE 2 To examine the robustness it is advisable to “abuse” the method in a controlled way.

2.38
secondary validation
demonstration by experiment that an established method functions according to its specifications in the user's hands

2.39
apparent selectivity
F
ratio of the number of target colonies to the total number of colonies in the same sample volume

$F = \lg(t/n)$

where

t is the apparent concentration of presumptive target types estimated by counting colonies;

n is the concentration of total colonies.

2.40
sensitivity
fraction of the total number of positive cultures or colonies correctly assigned in the presumptive inspection

2.41
specificity
fraction of the total number of negative cultures or colonies correctly assigned in the presumptive inspection

2.42
standard uncertainty
uncertainty of the result of a measurement expressed as a standard deviation

NOTE See reference [5].

2.43
type A evaluation
⟨of uncertainty⟩ method of evaluation of uncertainty by the statistical analysis of a series of observations

EXAMPLE Observations may be e.g. standard deviation, relative standard deviation.

NOTE 1 See references [5] and [6].

NOTE 2 Repeatability and reproducibility are often estimated by carrying out collaborative method performance tests where several laboratories study "identical" samples provided by a central organizer [15].

2.44
type B evaluation
⟨of uncertainty⟩ method of evaluation of uncertainty by means other than the statistical analysis of series of observations e.g. from assumed probability distributions based on experience or other information

NOTE See references [5] and [6].

2.45
uncertainty
⟨of measurement⟩ parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand

NOTE See reference [6].

2.46
uncertainty
⟨of counting⟩ relative standard deviation of results of repeated counting of the colonies or particles of the same plate(s) or field(s) under stipulated conditions

EXAMPLE Stipulated conditions may be e.g. the same person or different persons in one laboratory, or different laboratories.

2.47
validation range
range of mean number of particles per analytical portion for which obeyance of validation specifications (particularly linearity) have been acceptably demonstrated

NOTE It is usually expressed as the range of "reliable" colony counts.

3 Arrangement of the document

The first part (clauses 4 to 8) of this Technical Report contains informative material on basic principles, characteristics and limitations of microbiological methods, as well as on general aspects of validation. The second half (clauses 9 to 11) is the actual validation document, containing specifications and recommended procedures for their determination.

Old and new concepts and principles are not completely defined in the body of this Technical Report. Three annexes are attached. Annex A details the statistical formulae most relevant to this document, annex B contains numerical examples and annex C gives detailed plans for two validation experiments.

Statistical tests in the ordinary sense are not central to the ideas. Mathematical calculations are used mainly for the purpose of providing convenient summaries of data and statistical distributions provide guidance values. A table of the χ^2 distribution is the guide most frequently consulted.

The two BASIC programmes given in annex A are easily copied into desk-top computers or programmable pocket calculators to help with the basic calculations most frequently needed.

4 Basic concepts

4.1 General

As far as particle statistics is concerned, microscopic counts obey the same laws as viable counts but they are, with the exception of microcolony methods, free from the biological problems associated with growth. Differential stains, specifically labelled complexes or other agents used for finding the target do not change the metrological principles. The same validation principles as with selective colony methods can be applied.

Plaque counts of bacteriophages are in most respects similar to bacterial colony counts.

4.2 Validation

4.2.1 General

Validation means a process providing evidence that a method is capable of serving its intended purpose: to detect or quantify a specified microbe or microbial group with adequate precision and accuracy. The total count methods do not have a definable target group and can only be validated in relation to other methods or against theoretical expectations of precision.

Validation is classified as primary or secondary according to its purpose.

4.2.2 Primary validation

Primary validation is an exploratory process with the aim of establishing the operational limits and performance characteristics of a new, modified or otherwise inadequately characterized method. It should result in numerical and descriptive specifications for the performance and include a detailed and unambiguous description of the target of interest (positive colony, tube or plaque).

Primary validation characteristically proceeds by the use of specially designed test schemes.

A laboratory developing an in-house method or a variant of an existing standard should carry out the steps of primary validation.

It is imperative that technicians involved in primary validation have considerable experience with other microbiological methods.

4.2.3 Secondary validation

Secondary validation (also called verification) takes place when a laboratory proceeds to implement a method developed elsewhere. Secondary validation focuses on gathering evidence that the laboratory is able to meet the specifications established in primary validation. Presently, specifications are not available for most of the methods. Results of external quality assurance (see 4.2.8) may have to be used as the first step towards complete secondary validation.

Typically, secondary validation uses selected and simplified forms of the same procedures used in primary validation, but possibly extended over a longer time. Natural samples are the optimal test materials and the work need only address the procedure within the operational limits set by primary validation.

4.2.4 Analytical quality control (AQC)

Application of valid methods in their specified reliable limits does not automatically ensure valid results. Analytical quality control (AQC) used in connection with daily routine analyses is necessary. It controls the ability to use a method successfully.

AQC is a continuous process. Guidance charts, with limits derived from method specifications (from primary validation) or from theoretical considerations are the principal tools.

The methods of AQC are extensions of the routine analytical process, e.g. replications at different levels, or simply calculations not ordinarily performed on routine data. In addition, reference materials, intercalibrations and spiked samples are used.

Analytical quality control is needed in connection with primary and secondary validation. Only results reliable in the AQC sense should be used for derivation of validation criteria and performance characteristics.

International and national working groups have produced numerous documents on analytical quality control of microbiological methods (e.g. references [1, 2, 3, 4, 7, 8, 20, 21, 24, 26]). Standards manuals also contain sections on that subject. Although vital to validation, the methods of analytical quality control are not detailed in this Technical Report. In everything that follows, it is assumed that laboratories have the appropriate analytical controls and internal and external quality assurance systems in operation.

4.2.5 Equivalent methods

It is necessary to apply two methods in parallel on the same samples when developing an in-house method, and also when collecting information to justify the use of an alternative method.

Method performance consists of many aspects. There is no single test of method equivalence, nor numerical criteria for it. One method may be superior in specificity but inferior in recovery. All the collective information about robustness, precision and specificity gained during validation tests can be used for method comparisons (examples B.2, B.3, B.4 in annex B). The methods only need to be tested in parallel for recovery comparisons.

A method giving the highest recovery of confirmed target organisms is obviously the best, unless confirmation is always required for routine use. A method giving somewhat lower recovery but not requiring confirmation may be preferable. If high false negative rates or false positive rates observed in primary validation cannot be corrected by more refined target colony definitions, the method should be deemed invalid.

4.2.6 Test materials

It is a popular notion that validation should simulate routine as much as possible. Natural samples with natural concentrations of microbes should therefore be the main test materials. There are exceptions under some circumstances.

Artificial materials (certified reference materials and spiked samples) are used in internal and external quality assurance systems to ensure the basic proficiency of the laboratories participating in method validation excercises.

Spiking may be useful and even necessary in secondary validation whenever it is difficult to find natural samples with target organisms. Laboratory personnel will be able to familiarize themselves with the target.

Negative samples (blanks) should be limited to internal quality assurance. Their inclusion among samples studied for method equivalence may lead to a false impression of a good correlation between methods. If it were possible to know in advance which natural samples contain no target organisms, they would be a suitable selection for testing false positives in actual validation excercises.

The optimal concentration range for the validation of microbiological methods is narrower than the projected application range. High concentrations are unnecessary. Such samples resemble pure cultures and do not put the performance of the method or the laboratory to test.

Samples with very low bacterial content need to be studied for public health reasons, but are ill suited for method comparisons and other validation exercises for statistical reasons. The problem is mostly avoidable, because microbiological methods are generally not concentration-sensitive at the low end of the scale. Each individual germ reacts with the nutrient medium almost independently of other particles in the sample. If a method has a low recovery compared to another, the fact is more readily discovered with twenty or thirty colony-forming particles per plate than with one or a few.

Methods found valid at concentrations sufficient for validation are trusted to work also at low analyte concentrations.

4.2.7 Samples — Representativeness and sufficiency

Statistical theory provides solutions for calculating the number of samples required for different testing or estimation situations [3, 13]. To be able to make use of the theory, the size of real effects of importance and the power for their detection should be defined. An estimate of the uncertainty (precision) of the determination should be available and random sampling should be practiced.

Many or all of the above requirements are difficult to meet in advance planning and execution of microbiological method performance tests. Statistical techniques, if used at all, become rough guidelines.

The number and variety of samples examined ought to be sufficient to be convincing. Without the help of statistics, there are no exact ways of deciding. In some instances, the first sample studied might give the answer that the method is not good enough. Usually, however, more samples are needed. It may take a thousand samples to "prove" that two P/A methods are not equivalent. Choice of too few examples may be a waste of time.

4.2.8 External quality assurance and other collaborative tests

Participation of several laboratories in studies of "homogeneous" material are considered essential tests of both method and laboratory performance. (After outliers have been recognized and deleted, the remaining data are thought to provide the necessary information on method performance and proficiency.)

Collaborative tests have been developed into a tool widely applied for testing precision characteristics of chemical methods [34]. It seems somewhat premature to fully recommend the same in microbiology. It is assumed that all the participating laboratories have several years of experience with the methods tested and a proven ability to use them. The present experience is that collaborative experiments intended for method performance testing tend to turn into laboratory proficiency tests and training excercises.

A number of microbiological methods have been in use for decades (e.g. Endo agar for total coliforms, mFC for thermotolerant coliforms, m-*Enterococcus* agar for intestinal enterococci) by hundreds of laboratories. These methods would therefore theoretically be suitable objects for collaborative method performance testing.

When making collaborative proficiency tests for specific target organisms with selective media, the samples almost necessarily should be spiked with pure cultures or mixtures of organisms. Another solution is to use certified reference materials. This is a simplified and artificial situation. Major difficulties experienced by different laboratories in the routine use of methods on natural samples may be missed. As long as the detailed performance characteristics of a microbiological method have not been expressed quantitatively, these types of external quality assurance (EQA) schemes may nevertheless be the most satisfactory means towards secondary validation (verification) of a method.

4.3 Detectors

4.3.1 General

It is often convenient to call the nutrient medium in its container a detector (2.11). Two types of detector, solid and liquid, are employed in different microbiological method variants. They are also mostly associated with different enumeration or detection principles: liquid with P/A and MPN, and solid with colony counts.

All forms of validation in microbiology focus on the performance of the detectors.

The set of tubes (MPN) or the series of (countable) plates used for analysis is called a detection set or a detector set (2.12). Each individual MPN tube is a P/A detector.

EXAMPLE An individual well of a microtitre plate is a P/A detector. The whole plate when used as an MPN system is the detection set.

4.3.2 Detector comparisons

For most colony-count methods a liquid counterpart with the same chemical composition but without the solid matrix (agar, membrane filter) can be produced. The effect of the solid environment can be evaluated by comparing colony counts with the equivalent MPN estimate provided that the reaction for target recognition is the same on both types of detectors and the number of parallel tubes is large enough for adequate precision. Also the sensitivity of P/A detectors can be evaluated by similar liquid-solid comparisons.

4.4 Performance characteristics

Performance characteristics should be quantifiable and testable to be of use in validation.

The terminology on performance characteristics in this Technical Report mostly follows the chemometric usage. Because the original definitions of the terms do not always fit microbiological methods perfectly, they have been modified and adapted as necessary.

The performance characteristics dealt with in this Technical Report are related to scope (list of situations and sample types where the method is applicable), precision, linearity, recovery, working limits in terms of lowest and highest recommendable colony number per plate, selectivity, specificity and robustness (ruggedness). Definitions of these and other terms can be found in clause 2.

4.5 Specifications

Specifications are either numerical or qualitative expressions of performance characteristics or of working limits derived from them. Primary validation should provide the following:

a) morphological identification of the (presumptive) target;

b) statements regarding incubation conditions (temperature, time, gas atmosphere, moisture) and media characteristics (pH, stability);

c) a statement regarding reliable working limits in terms of colony or plaque numbers per detector (plate, membrane filter) if possible;

d) expressions of uncertainty within the specified reliable limits;

e) scope and limitations.

5 Limitations and characteristic features of microbiological methods

5.1 Recovery of the analyte

The microbiological analyte consists of discrete living particles, variously called colony-forming units (CFU), colony-forming particles (CFP), germs, propagules, etc. (see clause 2). The number of colonies observed is an approximation of the number of living particles.

The useful arsenal of performance tests is limited by the near impossibility of knowing the true amount of the analyte in a sample or in an analytical portion. Detectors cannot be challenged with an exactly known number of germs.

Viability is defined by growth, i.e. by the method itself. Absolute recovery is undefinable and traceability is impossible. As viability may appear different with different detectors and/or with different sample history, the relative recovery (involving the new method and a reference) is a practicable performance characteristic even though the true result is not known.

5.2 Sample variance

In the environment and even in laboratory samples, the distribution of particles is uneven. The sampling variance of the environment is not a characteristic of the method, whereas the subsampling variance of a laboratory sample may be considered so. It may not be possible to use mixing practices sufficient to ensure perfect mixing of sample contents without some loss of viable cells. Within-sample variance often remains considerable and causes problems in validation. Performance, especially precision and upper working limits, will need to be determined separately for different matrices.

5.3 Particle distribution and overdispersion

Random variation due to uneven distribution of particles between parallel samples, even in perfectly mixed suspensions, is a characteristic feature of microbiological methods [31].

The basic random variation is unavoidable and has nothing to do with technical skills or equipment. It follows a known mathematical law, the Poisson distribution [28], and is therefore accountable.

Technical imperfections and many other causes are responsible for additional variation. Parallel determinations vary even more than is explained by the Poisson distribution. This situation is called overdispersion (2.20)

Many arguments support the negative binomial distribution (2.19) as an overdispersion model in microbiology [11,14,15].

5.4 Interactions in the detector

Considerable difficulties arise from interactions of abiotic and biotic factors within the detectors. Liquid detectors suffer from possible cohabitation of a varied microflora in the same tube. Colony detectors suffer from crowding and masking of target colonies by debris and non-target growth. Reading becomes difficult, unreliable or impossible.

5.5 Robustness

Microbiological methods are not robust. Both the analyte and most of the impurities are living. This may cause unexpected phenomena on the detectors. Microbiologists should be able to recognize and understand them. Matrix effects, incubation stress, quality and origin of substrate ingredients, composition of the sample microbial flora, and training of the technicians can all affect the result.

The origins of the lack of robustness are in five principal areas: the sample (its physicochemical properties and microbial population), the competence of the personnel, sample preparation, incubation conditions, and the detection medium. Of these, the last three should be standardized.

Primary validation should establish the general limits within which methods are expected to perform well. The efficient statistical design for robustness testing developed by AOAC [34] can be applied in microbiology as well.

Considering the living nature of the analyte, the freedom of choice as regards reaction times (incubation time) and incubation temperature is often surprising. Colony counts are more time-sensitive than normally assumed. For example, International Standards, in their effort to include practices current in different countries, frequently imply robustness within improbable limits.

EXAMPLE The method description for detection and enumeration of coliform organisms allows incubation for 18 h to 24 h at 36 °C ± 1 °C. This implies considerable stability of results towards variations in incubation time and temperature; probably more than is actually true.

When temperature is one of the selective factors (e.g. 44 °C with thermotolerant coliforms), even the position of the plate in the incubator or in the stack of plates can have strong effects on the result. This has been recognized in some standards by specifying the highest permissible stack size (e.g. six). Total elimination of the stacking effect would require incubation in one layer only, which is considered impracticable.

Another important robustness feature that is seldom considered is the storage of samples before analysis. It is believed to be generally valid that samples can tolerate refrigerated storage for e.g. 24 h [9]. The cold-stress inflicted during refrigerated storage may be unimportant for total counts but harmful for methods depending on highly selective methods. Thermal shocks and other additional stresses are not general but method-specific.

5.6 Spurious errors

A typical feature of most microbiological colony-count methods is their predisposition to unexpected problems, often on a single plate. The different techniques (plating, surface spreading, MF) are quite unequal in this respect. The problems may be due to the presence of a single disturbing colony, moisture, temperature differences, solids and other impurities in the test portion, contamination, etc. Such "accidents" do not reflect the method performance in general and are neither quantifiable in validation nor predictable by mathematical modelling. When very frequent they make the use of a method uncertain. Liquid culture methods are less susceptible.

The "state of the art" adherents hold the view that spurious errors belong to microbiological methods and should be included in the uncertainty estimate of the performance. This is related to the idea that methods ought to be validated as they appear in use.

The view taken in this Technical Report is that spurious errors are the concern of daily AQC. Although their detection depends on the competence of the technicians, every effort should be made to recognize the erroneous values and delete them. If spurious errors are found to cause frequent failure of analyses, the method is obviously unsuitable for the intended purpose.

5.7 Control and guidance charts

An analytical procedure is a type of a process, and the resulting measurement or determination can be considered its product. The basic control-chart idea, born in the process industry, might seem to work in the control of the analytical process as well. It would help detect whether sudden or more insidious changes in process quality have taken place. When an out-of-control situation is more or less certain, the process can be stopped and repaired.

The possibilities of process control in analysis are limited. The most common types are the use of reference samples to test constancy of recovery (absence of systematic bias) and duplicate determinations to check precision (repeatability and/or reproducibility).

This thinking suits automated chemical analyses quite well, but is out of place in microbiological analyses. It assumes that the process (i.e. analytical procedure) is at its best performance in the beginning and may get worse suddenly or gradually, whereafter all results are in error. In microbiology, it is not easy to point out when a method is in or out of control. An analytical result of poor quality may not be related to any other results of the same day.

Graphical plots are useful and practical devices not only in AQC but also in certain validation contexts. For reasons indicated above, their use in microbiology should be limited to give guidance on the analytical practice. They should not be used for condemning the analytical work of a whole day or of a series of results.

It has been suggested that the term "control chart" should be restricted to obvious control situations, such as monitoring incubator temperatures. The term "guidance charts" is preferred when illustrating graphically the performance of microbiological methods. Some guide values may nonetheless be chosen to help in decision-taking.

6 Mathematical models of variation

6.1 Unavoidable basic variation — The Poisson distribution

6.1.1 General

The chance variation of particle numbers between parallel test portions at the optimal working range of microbiological detectors is considerable even if the suspension is perfectly mixed (completely random) and no technical uncertainties of measurement are involved.

This creates a typically microbiological dilemma. The colony-count methods are, in most biological and technical respects, at their best when numbers are so small that statistical precision is inadequate.

6.1.2 Precision of colony-count detectors

Precision is usually expressed in terms of the standard deviation. The basic statistical variation of counts can be mathematically modelled by the Poisson distribution.

The variance (s^2) of the Poisson distribution is numerically equal to the mean (m). (Equality of mean and variance does not prove that the data follow a Poisson distribution.) The relative standard deviation (RSD) is in inverse relation to the mean count or more generally to the total count (c) of the detection set:

$$\text{RSD} = \frac{s}{c} = \frac{\sqrt{c}}{c} = \sqrt{\frac{1}{c}} \qquad (1)$$

EXAMPLE A single plate count from a sample of a perfectly mixed suspension, say 48 colonies, has the theoretical relative precision (RSD) of $1/\sqrt{48} = \pm 0{,}14$ (CV = 14 %). If the same total count, 48 colonies, had resulted from three parallel plates 12 + 16 + 20 = 48, the relative standard deviation of the mean 48/3 = 16 would be the same.

The dependence of the relative precision (CV, coefficient of variation) on the particle count (c) is shown in Figure 1.

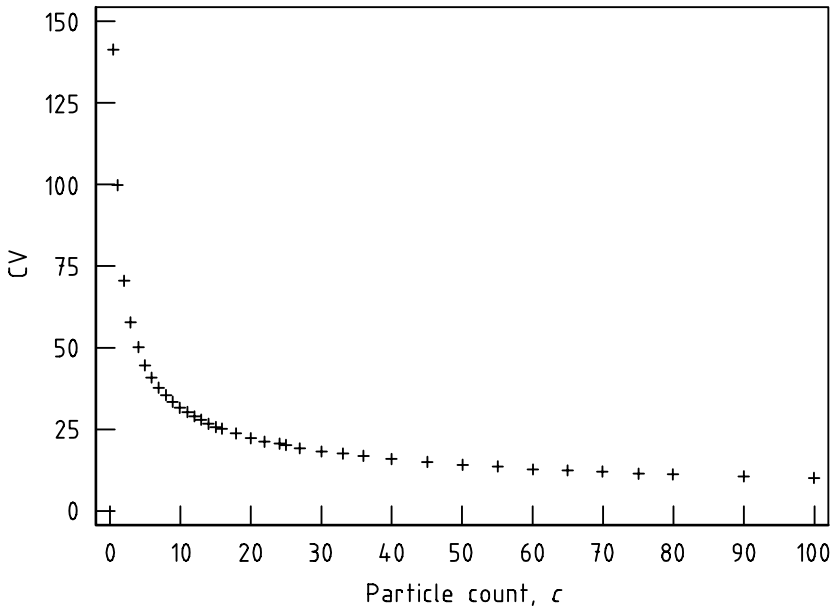


Figure 1 — The coefficient of variation of colony number c in a perfectly mixed suspension following the Poisson distribution law

The graph shows why colony numbers such as 20, 25 or 30 have been traditionally considered the lowest statistically reliable counts.

The Poisson model can be used for estimating the theoretical statistical uncertainty at any colony count and conversely for making decisions on the lower working limits based on stipulated statistical precision.

Random uncertainty increases rapidly as the colony count decreases (Figure 1). In the count range below about ten, which happens to be of considerable public health interest, single measurements are so imprecise that they can hardly be characterized as better than semi-quantitative.

The count range 1 to 10 approximately corresponds to the range that determines the characteristic precision of the common 3×5 tube MPN estimate. Classifying colony counts below ten as semi-quantitative is in harmony with the opinion held by many statisticians that MPN methods according to the common 3×3 or 3×5 designs are more appropriately considered as tests of the order of magnitude rather than quantitative estimates. (The 3×5-tube MPN estimate has the 95 % confidence range of about $\pm$ 0,5 logarithmic units [12].)

6.1.3 Precision of the MPN detectors

The precision of MPN estimates depends on the count itself, but in a somewhat more complicated way than that of the colony count. The standard deviation of logMPN is a wavy curve. It has as many local minima as there are dilution levels in the detection set. As an example, the logarithmic standard deviation was calculated for the 3×10 tube MPN detector with the help of a computer program [19] (Figure 2).

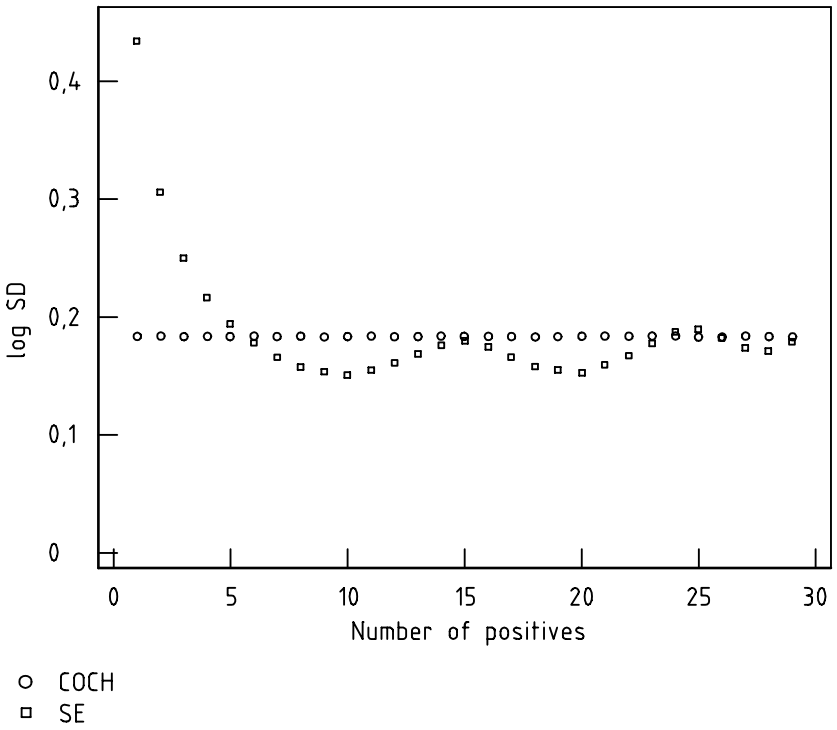


Figure 2 — The logarithmic standard deviation of 3 × 10 tube MPN

The number of positive tubes in the whole set is on the abscissa. The undulating curve displays the true standard deviation. The straight horizontal is Cochran's [12] approximation.

Cochran [12] proposed an approximate constant standard deviation for the entire MPN range. This approximation is indicated by the horizontal line in Figure 2. It is seen that the exact result deviates most from the approximation when most of the tubes in the detection set are negative.

According to Cochran's [12] formula, precision of the MPN estimate in logarithmic scale depends in a simple way on the number of tubes (n) and on the dilution factor (f) between consecutive dilutions. The constant 0,58 was chosen "by the eye" for tenfold dilutions to give the fit illustrated in Figure 2. If the dilution factor between dilutions is less than ten, then the constant 0,55 can be used [12].

$$\text{SD of lg MPN} = 0{,}58\sqrt{\frac{\lg f}{n}} \tag{2}$$

The 95 % confidence limits in many MPN tables are based on Cochran's approximation in equation (2), but are being replaced by the more exact "true" limits in recently published tables.

The standard deviation of any individual MPN value is nowadays easily obtained by an appropriate computer program, e.g. [19].

Despite its approximate nature, Cochran's formula is useful in experimental planning and method comparisons (see example below).

EXAMPLE Assume a determination based on the MPN detection system of 3 × 32 wells in a 96-well microtitre plate. Suppose further a dilution factor f = 3 between consecutive dilutions. The standard deviation of lg MPN, according to Cochran's approximate formula, is

$$\text{SD of lg MPN} = 0{,}55\sqrt{\frac{\lg 3}{32}} = 0{,}55\sqrt{0{,}014\,9} = 0{,}067\,2 \tag{3}$$

With high numbers of parallel tubes and with dilution factors less than 10, the MPN systems become equal or better than the colony methods because of the superior growth conditions, ease of interpretation and lack of overlap errors.

6.1.4 Limit of detection — Poisson model

At very low particle concentrations all microbiological methods, MPN and colony count included, become essentially P/A methods. There is little harm, as far as public health or product quality assessment is concerned, in considering any counts below three or four as mere positive detections of presence.

If the limit of detection is defined in terms of the probability of scoring a positive result, it cannot be read from the graph in Figure 1. The probability of a positive result $p(+)$ when the Poisson distribution prevails can be calculated from

$$p(+) = 1 - e^{-m} \tag{4}$$

where

e is the base of natural logarithms;

m is the mean number of particles per analytical portion.

EXAMPLE One of the popular definitions for the limit of detection is the concentration at which the probability of detecting the presence of the analyte equals 95 % [$p(+) = 0,95$].

If $p(+)$ is taken as 0,95, then $e^{-m} = 1 - 0,95 = 0,05$. Solving the equation for m yields $m = -\ln(0,05) = 3,0$ Thus, at the average count of 3, the chances of detecting a particle in a test portion equals 0,95 (provided that the Poisson distribution prevails).

6.2 Overdispersion — The negative binomial model

6.2.1 General

Preparation of the basic suspension, dilution, inoculation and counting of the colonies are not totally free of uncertainty. Every technical step adds to the total variability of the measurement. Parallel determinations involving the whole analytical procedure cannot be expected to follow the Poisson distribution. Overdispersion, i.e. variation greater than fully random (in the Poisson sense), between parallel determinations will be observed.

Overdispersion is the normal state of microbiological determinations and Poisson distribution is an exception.

The common causes of overdispersion, apart from the spurious errors, have effects roughly proportional to the mean or actually to the total number of colonies (c) counted in the detection set. Other arguments why the overdispersion of microbiological counts should follow this pattern have been presented elsewhere [11,14]. As the basic uncertainty due to the random scatter of particles in suspension follows the Poisson distribution the total variance s^2 can be written as

$$s^2 = c + u^2 c^2 \tag{5}$$

where

u is the overdispersion factor, relative standard deviation of addional uncertainty;

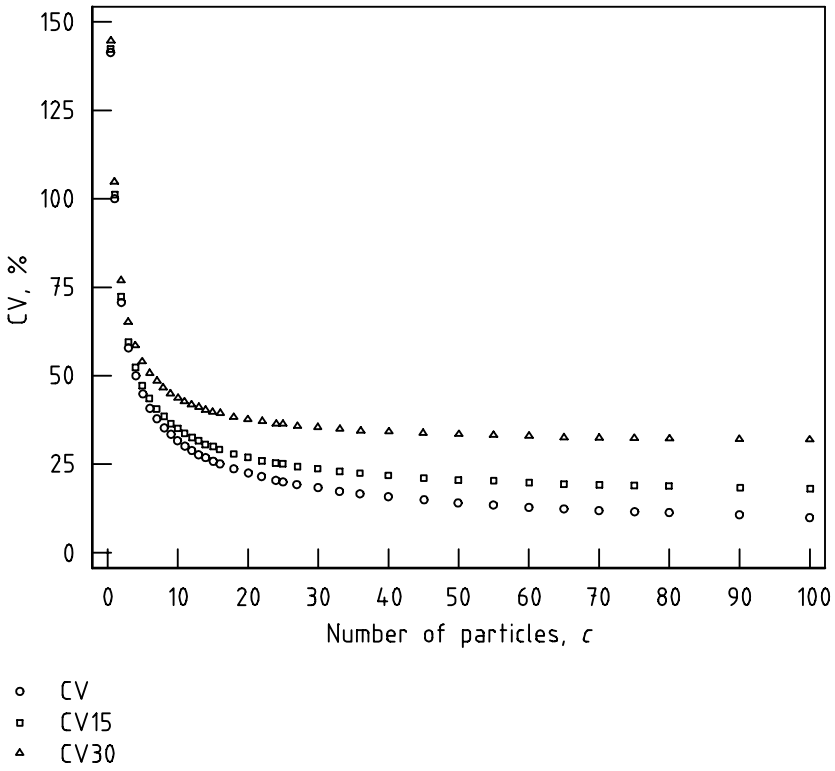
c is the colony count in the entire detection set.

The first part of the variance (c) is due to the Poisson process, the rest (u^2c^2) is due to the combined effect of all the random overdispersion factors. A statistical distribution with this model of variance is called a *negative binomial distribution* (also heterogeneous or compound Poisson distribution).

Consequently, the relative standard deviation can be expressed as

$$\mathrm{RSD} = \sqrt{\frac{1}{c} + u^2} \tag{6}$$

Figure 3 shows the effect of different degrees of overdispersion on the total relative precision (coefficient of variation).



NOTE Dependence of the coefficient of variation (CV) on the number (c) of particles counted and moderate additional variation. Lower curve: the Poisson model with no overdispersion. Upper curves: negative binomial with 15 % and 30 % overdispersion (u = 0,15 and 0,30).

Figure 3 — Effect of overdispersion on CV

The colony number required to reach a given total relative precision is considerably higher in an overdispersed situation than in the totally random (Poisson) case. It can be calculated by solving equation (6) for the colony number c:

$$c = \frac{1}{\mathrm{RSD}^2 - u^2} \tag{7}$$

EXAMPLE To achieve the relative standard deviation RSD = 0,2 when the overdispersion is u = 0,15 requires, according to equation (7), the colony number $c = 1/(0{,}2^2 - 0{,}15^2) = 1/(0{,}04 - 0{,}0225) = 57$. The same precision is reached in a fully random (Poisson) situation with the colony number $c = 1/0{,}2^2 = 1/0{,}04 = 25$.

NOTE It is obvious that total precision lower than the overdispersion cannot be achieved within a single determination. The whole procedure must be repeated if better precision is required. With n parallel determinations, the total relative standard deviation can be roughly estimated from

$$\mathrm{RSD} = \sqrt{\frac{1}{\sum c} + \frac{u^2}{n}} \tag{8}$$

where

$\sum c$ is the total number of colonies recorded;

u is the overdispersion constant;

n is the number of parallel determinations.

6.2.2 Limit of detection — Negative binomial model

The detection limit, if defined in terms of the probability of a positive result can be calculated from the probability of negatives. Quoting Anscombe [11] but changing the symbols to the ones used in this Technical Report, the probability of a negative result (probability of zero) is given by the formula:

$$p_0 = \left(1 + u^2\, c\right)^{-1/u^2} \tag{9}$$

Solving for c gives the detection limit when the probability of negatives and the overdispersion factor have been given.

$$c = \frac{p_0^{-u^2} - 1}{u^2} \tag{10}$$

As is evident from Figure 2, the detection limit is rather little affected by moderate overdispersion (see example below).

EXAMPLE The mean colony count required in order to achieve a 95 % probability of a positive result in an overdispersed situation depends on the overdispersion factor. Assume an overdispersion factor u = 0,30. Direct substitution of the probability of a negative result $p_0 = 1 - p(+) = 1 - 0{,}95 = 0{,}05$ in equation (10) yields $c = (0{,}05^{-u2} - 1)/u^2 = (0{,}05^{-0{,}09} - 1)/0{,}09 = 3{,}44$. The corresponding estimate with the Poisson distribution (no overdispersion) would be $c = \ln(1/0{,}05) = 3{,}00$ (see example in 6.2.4).

6.2.3 Quantifying overdispersion

There are three major ways to estimate the parameter u [20].

In the range of mean values and overdispersion factors that are of interest in the validation of colony-counting methods, Anscombe's Method I [11] is efficient. It consists of solving equation (5) for u.

To be able to use Anscombe's Method I, a substantial series (preferably more than 30) of independent observations on a single sample must be available to base reliable estimation of the variance (s^2) and mean (m) on. Equation (5) yields, when substituting m for c

$$u^2 = \frac{s^2 - m}{m^2} \tag{11}$$

To be efficient, the observations should be limited to colony counts in the optimal working range. The mean should never be smaller than 30. This applies especially if the purpose is to estimate u in a single experiment.

EXAMPLE To demonstrate the feasibility of overdispersion calculations, an experiment was made to deliberately introduce overdispersion in a series of parallel counts. 24 membrane filtrations were made from a carefully mixed suspension of a pure culture of *Enterococcus faecium* in such a way that eight plates each were inoculated with 8 ml, 10 ml and 12 ml of the suspension. This introduced a volume imprecision which was bound to cause a roughly predictable amount of overdispersion to the 24 otherwise parallel counts.

The volume calculated from 8 × 8 ml, 8 × 10 ml, and 8 × 12 ml has a mean μ = 10 and standard deviation σ = 1,633. The relative standard uncertainty due to overdispersion should therefore be $u = s/m$ = 0,163 3 (CV = 16,33 %), if the volumes were exactly as stated.

The 24 colony counts observed had a mean $m = 379{,}29$ and variance $s^2 = 4\,586{,}54$. According to equation (11), $u^2 = (4\,586{,}54 - 379{,}29)/379{,}29^2 = 0{,}029\,2$. Hence, the overdispersion coefficient $u = \sqrt{0{,}029\,2} = 0{,}171\,0$ which is close enough to the expected value, considering that it also includes the possible counting and volume uncertainties ignored in the expected $u = 0{,}163\,3$. (All the 8 ml, 10 ml and 12 ml measurements were believed to be exact.)

The uncertainty estimate calculated from the above is not generally valid, as it is based on one sample only.

The second solution is also based on equation (5) but considers several samples. Dividing the equation by c yields

$$Y = \frac{s^2}{c} = 1 + u^2 c \qquad (12)$$

Whenever parallel determinations are available, estimates of variance and mean can be calculated. Computing the ratio $Y = s^2/c$ from each set provides a large number of (Y,c) pairs. The slope of the regression line fitted to the points gives an estimate of u^2. The random scatter is inevitably considerable if the estimates of mean and variance are based on small samples (small numbers of parallel determinations)(see example B.7). The advantage of this approach is that the estimate of overdispersion is based on a large selection of different samples and has therefore considerable general utility.

NOTE The above reasoning applies best when the mean (m) is the untransformed colony (or particle) number per detection set and the detection sets are in other respects identical in all parallel determinations.

6.2.4 Overdispersion at detector level

Parallel counts from a single suspension may also vary more than the Poisson distribution allows. This is a classical observation [16]. At this level the only causes of overdispersion are pipetting errors, counting uncertainty and spurious errors ("accidents"). Overdispersion is a useful measure of overall reliability. It can be detected by the indices of dispersion (X^2, G^2) (see example B.6 in annex B).

6.3 Statistical and practical limits

6.3.1 General

The lower working limits of microbiological methods are to a large extent matters of definition.

The upper limits are defined by the space requirements and ensuing interactions of microbial colonies. A microbial particle needs to multiply about a millionfold to become detectable by the naked eye.

6.3.2 Lower limit

When only one analytical portion (one plate) is examined, the statistical lower limit can be conveniently expressed as the lowest "reliable" count per plate; reliability being defined by the choice of precision.

All things considered, the lower limit should be chosen somewhere in the vicinity of 20 colonies per detection set. At that colony number the coefficient of variation in a completely random (Poisson) situation is ca. $\pm$ 25 % (6.1.4). If overdispersion is numerically known, calculation of the lower limit of determination can be based on the negative binomial model (6.2.2).

6.3.3 Upper limits

P/A analyses do not have an upper limit. As the number of germs in the analytical portion increases, the probability of detecting their presence approaches certainty.

With MPN methods, the practical upper limit has been exceeded when all tubes of all dilutions are found positive. These mishaps do not reflect the upper limit of the method itself, only mistaken dilution. No upper limit can be set for statistical reasons either, because precision does not depend in a simple direct way on the number of particles introduced in the detection set. It is characteristic of the MPN procedures that their precision can be improved at will by changing the configuration of the detection set (see 6.1).

With colony-count methods, precision theoretically improves steadily with the number of target colonies observed in the detection set. In practice, colony-count methods have an upper limit per detector which varies with the testing situation. The colony-count detector (agar plate, membrane filter) becomes "clogged" or saturated for several reasons, of which the number of target colonies is only one.

There are many causes and accordingly many ways of determining the upper limit. One possibility is to determine the colony count per plate where total uncertainty (systematic errors included) again rises to the same level as at the lower limit [30].

Another consideration is the crowding-out or masking of target colonies by non-target growth. Such a phenomenon can be quantitatively tied to the selectivity of the method, or more scientifically to the "coverage", i.e. fraction of available space occupied by colonies. Even pure cultures have an upper limit depending on the coverage [25].

A third aspect is loss of proportionality (loss of linearity) that occurs at high congestion of the plate.

6.4 General tests for randomness — Detection of overdispersion

Deviation from randomness (presence of over- or underdispersion) can be effectively detected by one of the two classical goodness-of-fit tests, viz. the Poisson index of dispersion (X^2, D^2) or the log-likelihood ratio statistic (G^2). Calculation of the indices is presented in annex A.

In the few instances (e.g. parallel plates) where gross overdispersion has no acceptable technical reasons, agreement with the Poisson distribution can be used as a test of method or analyst performance [16, 33]. Due to its simplicity this characteristic is especially useful as an AQC tool.

7 Specifications — Current practice

An example of how the best of standard protocols currently express method performance characteristics can be found in the way the *Standard Methods for the Analysis of Water and Wastewater* [8] instruct the users of their methods. In different parts of the total coliform membrane filter procedure, the user will find the following statements related to specifications.

a) Scope

The scope of the method can be inferred from a table giving recommended sample volumes for different types of waters. This table includes all types of drinking and recreational waters as well as wastewaters.

b) Incubation robustness and time-sensitivity

The main rule is given as follows: "Incubate for 20 h to 22 h at (35 ± 0,5) °C." In another part of the document exceptions to the main rule are presented: "Samples of disinfected waters may include stressed organisms that grow relatively slowly and produce maximum sheen in 22 h to 24 h. Organisms from undisturbed sources may produce sheen at 16 h to 18 h, and the sheen subsequently may fade after 24 h to 30 h."

c) Reliable working limits

The reliable working limits can be inferred from: "Compute the count, using membrane filters with 20 to 80 colonies..." This is followed by further reservations that tie the working limit with selectivity: "...and not more than 200 colonies of all types per membrane." Some reasons for the latter are found in another part of the document: "Size of sample will be governed by expected bacterial density, which in drinking water samples will be limited only by the degree of turbidity or by the noncoliform growth on the medium."

d) Target definition and identification

"All bacteria that produce a red colony with a metallic sheen within 24 h incubation at 35 °C on an Endo-type medium are considered members of the coliform group. The sheen may cover the entire colony or may appear only in a central area or on the periphery."

There are further considerations: "Endo-type media occasionally may produce dark red atypical colonies without a metallic sheen. Verification of various types of typical (sheen) and atypical (non-sheen) colonies will detect false negative results and provide experience in colony recognition."

e) Other limitations and specifications

The standard document also offers advice and makes demands on quality control of media and equipment.

8 Specifications — Recommended approach

Generally speaking standards currently provide little help for laboratories seeking to make sure that they apply the methods well and obtain valid results.

What seems to be lacking is a concise presentation of what laboratories should do to verify that the method also works in their hands properly and how to recognize good from bad performance.

A section of specifications should be added to all written standards.

The format for colony count methods might include the following:

a) Sensitivity**:** with few exceptions (cases mentioned) more than 90 % of presumptive positives are confirmed.

b) Selectivity: usually better than –1 (see definition of selectivity). Results are not valid if selectivity is less than –2.

c) Counting uncertainty: the relative standard deviation of duplicate (replicate) counting within a laboratory is generally less than u_Z = 0,05. Individual counting uncertainty (one person) remains normally below $u_Z = \pm$ 0,03.

d) Parallel plating: variation is within the Poisson distribution. (If not, the extent of overdispersion should be given.)

e) The within-sample variation has a procedural added uncertainty coefficient of less than $u_X = \pm$ 0,10 in water samples. In solid samples, the added uncertainty should remain below 0,15.

f) Proportionality (linearity) of the detector depends on the selectivity. It is adequate up to the limiting colony numbers given below:

Selectivity	**Upper limit of linearity** (target colonies per plate)
0	500 (pure cultures)
– 0,5 to –1	200 to 100
– 1 to – 2	100 to 25
below – 2	P/A detection only

With excessive colony growth, the upper functional limit may be reached sooner. Irrespective of the value of selectivity, counts are not valid if more than 1/3 of the available space is occupied by growth (target and non-target)

The above is only an example of how concrete specifications of method performance might be presented. Such specifications are testable with appropriate designs. The numerical values are given as illustrations. They are not to be understood as standard values at present.

If information on relative recovery compared with a standard reference is available, it should also be given.

If reproducibility data from collaborative method performance tests are available, they should be added as further information.

In addition, testing conditions, target description and sample storage should be specified.

9 Determination and expression of performance characteristics

9.1 General

Performance characteristics based on frequencies of items in qualitative categories are called categorical in the following.

In connection with selective methods, they are determined by verifying presumptive positive and negative cultures.

9.2 Categorical characteristics related to specificity and selectivity

9.2.1 General

The most comprehensive qualitative view of method performance is gained by studying natural samples. The samples should cover the complete scope of the method, including different samples and pollution situations as well as different seasons.

A colony or a plaque is the equivalent of one P/A test or one tube in an MPN series. In primary validation both presumptive positive and presumptive negative cultures should be verified.

The most cost-effective and biologically attractive way of testing liquid culture methods (both P/A and MPN) is to use an MPN design. The tubes in dilutions where both positives and negatives occur are selected for verification. These are the dilutions where positive tubes arise from one or very few cells.

The performance characteristics associated with selectivity and specificity can be defined numerically [17]. They relate to the relative proportions of colonies or tubes assumed positive or negative on the basis of the first impression (presumptive) compared with the "truth" after verification. After n verification tests have been made, their results are divided into four categories:

a) number of presumptive positives found positive (true positives);

b) number of presumptive negatives found positive (false negatives);

c) number of presumptive positives found negative (false positives);

d) number of presumptive negatives found negative (true negatives).

These frequencies can be conveniently expressed in a 2 x 2 table:

		Presumptive count		
		+	–	
Confirmed count	+	a	b	$a + b$
	–	c	d	$c + d$
		$a + c$	$b + d$	n

The performance characteristics calculated from these observations are defined as follows:

1) sensitivity = $a/(a + b)$, the fraction of the total positives correctly assigned in the presumptive count;

2) specificity = $d/(c + d)$, the fraction of the total negatives correctly assigned in the presumptive count;

3) false positive rate = $c/(a + c)$, the fraction of the observed positives wrongly assigned;

4) false negative rate = $b/(b + d)$, the fraction of the observed negatives wrongly assigned;

The total number of tests is $a + b + c + d = n$.

5) efficiency E is a general single parameter, which gives the fraction of colonies or tubes correctly assigned:

$E = (a + d)/n$.

The colonies should be picked randomly from all colonies (target and non-target) considered as a whole. With liquid cultures, all positives and negatives are likely to be tested in their actual proportions.

Due to the strong influences of the operator and the microbial population, none of the performance characteristics detailed above can have a method-specific constant value.

Secondary validation need only be concerned with false positives, unless a sensitivity rate lower than specified seems recurrent.

9.2.2 Selectivity

Selectivity of a microbiological method is defined in this Technical Report as the logarithm of the fraction of presumptive target colonies (presumptive positives) among the total:

$F = \lg [(a + c)/n]$.

Selectivity can be so low in some types of sample or during certain seasons that a reliable target count is unobtainable. The upper working range is greatly influenced by selectivity.

The "presumptive" selectivity defined above is a tool selected for convenience. If resources are available, the real selectivity based on verified counts of true positives is scientifically better. Economic constraints may limit its use to final comparisons between methods.

Apparent selectivity varies considerably because of seasonal effects and other causes that change the relative abundance of the target and background populations. It is not a method-specific constant. Nevertheless, selectivity provides information about method performance and may help choose between alternatives. A wide variety of cases should be studied. Because no decision criteria are available, there are no grounds for deciding what specific number is sufficient. Enough data to be convincing one way or another should be gathered.

9.3 Working limits

9.3.1 Lower limits of detection and determination

The lower working limit is a matter of definition in all methods (see clause 6). Numerically, the limit should be expressed as the lowest sufficient number of colonies or particles per detection set or parallel plates.

9.3.2 Upper limits

Every method, in fact every individual determination, has an upper reliable limit. It is not a clearly fixed number, but a somewhat vague region of colony numbers where counts per plate become too uncertain to base a valid determination on.

The upper working limit of a colony detector is signalled by the following.

a) Overdispersion of parallel colony counts may become more pronounced and more frequent. The situation can be recognized by the help of the Poisson index of dispersion (Example B.3).

b) Colony counts from different dilutions (volumes) disagree. Proportionality (linearity) is lost. Proportionality can be tested by methods based on the G^2 index (annex A, and Example B.4).

The proportionality (linearity) test is the more effective of the two. Both features can be explored by the use of a single specially designed proportionality test (annex C) [27, 33] or by collecting data on the above elements from separate tests.

9.4 Working range of MPN procedures

The lower and upper practical limits of MPN procedures are the extreme cases when only one tube in the detection set is positive or negative.

The apparent application range of a 3×5 detection set is accordingly 0,2 to 160 particles per unit volume (analytical portion) of the first (least diluted) series, irrespective of the nutrient medium or target population.

9.5 Precision

9.5.1 General

Each analytical measurement should be accompanied with a precision estimate, even though it is not feasible to determine experimentally the precision of each individual determination. There is, therefore, considerable interest in deriving generally valid precision statements for different methods.

Ideally, primary validation should provide a general estimate on which the precision of any measurement can be based. this is only possible in microbiology in two ways. One must either assume total randomness, and consequently the validity of the Poisson distribution, or determine a general constant of overdispersion by experiment.

Precision estimates can be obtained in two principally different ways that have been named Type A and Type B (see clause 2) [5, 6].

9.5.2 Type A precision estimates

Type A estimates are derived from statistical calculations based on parallel determinations. They are expressed as standard deviation (standard uncertainty) or relative standard deviation.

The most general Type A estimates are obtained by collaborative method performance tests according to principles mainly developed by AOAC [18]. The Type A estimates of microbiological measurements so far have been calculated and expressed as standard deviation in the logarithmic scale.

Such estimates have not yet made their appearance in microbiological standards.

As is apparent from the discussion in clause 6, the precision of microbiological determinations is not constant even in logarithmic scale, but depends on the colony number. That is one reason why general Type A estimates are always approximate and tend to be large.

9.5.3 Type B precision estimates

Type B estimates are based on assumed probability distributions or other information.

The Poisson distribution has been considered an appropriate statistical model in perfectly random suspensions in microbiology. Precision statements based on the Poisson distribution are quite common in microbiological standards.

EXAMPLE The American Standard Methods includes the following Type B uncertainty statement: "For results with counts, c, greater than 20 organisms, calculate the approximate 95 % confidence limits using the following normal distribution equations:

Upper limit = $c + 2\sqrt{c}$

Lower limit = $c - 2\sqrt{c}$ "

The standard deviation in these cases is obviously based on the assumption of equality of variance and mean (Poisson).

For reasons discussed in clause 6, the Poisson model is highly idealized and yields minimal uncertainty estimates. It has the distinct advantage that an individual estimate can be attached to each determination.

9.5.4 A mixed model

An overdispersion model based on the negative binomial distribution combines the theoretical Poisson element with an empirical overdispersion constant (see 6.2.3).

Determination of the overdispersion factor is a realistic goal for primary validation. Methods to determine it are treated in (6.2.3).

The uncertainty value should be given for different applications of a method if it varies from situation to situation. It may depend on the matrix being studied. It is conveniently expressed as the relative standard deviation.

9.5.5 Precision of MPN

The MPN estimates rely on the assumption of complete randomness of particle distribution in the detection set. With the low particle numbers involved, this trust is not easily challenged by experimental testing.

The precision of MPN estimates is determined by the number of parallel tubes per dilution and by the dilution coefficient. An approximate formula, assuming a constant uncertainty over the range, is given in [12]. The 95 % confidence limits for the standard combinations of tubes are published in MPN tables (see also 6.1).

10 Procedures and steps of validation

10.1 General

Validation almost necessarily proceeds in steps. The actual method performance investigations should be preceded by a search of the reliable working range in which method performance is relevant. Any recovery comparisons between two methods should be limited to cases for which both methods are reliable.

10.2 Primary validation

10.2.1 Target identification

Pure culture studies during initial development of the method provide the basic description of the target colonies or P/A tubes. Obviously, more than one pure culture strain of the target shall be tested.

The validity of the basic description shall be proved by using the candidate method on a representative selection (see 4.2.6) of natural samples. Presumptive target colonies, plaques or cultures are tested by isolating cultures and making appropriate confirmation tests. Presumptive non-target colonies or negative tubes are tested in the same way.

Numerical values of sensitivity, selectivity, false positive and false negative rates and efficiency are calculated.

Target description is amended if necessary.

10.2.2 Counting uncertainty and time-sensitivity

After target morphology has been properly defined, the method is tentatively used on a collection of natural samples.

The baseline reliability of the counts is studied by repeated counting of the colonies of the same plates within a short time. If more than one person in the laboratory will be working with the method, they should all be involved.

Time-sensitivity of the results is conveniently studied in the same connection by counting the same plates at two points in time: at the beginning and end of the assumed incubation-time tolerance range. The effect is evaluated using the uncertainty of repeated counting as the basis of comparison.

The excercise should produce individual or collective numerical estimates of counting uncertainty, expressed as relative standard deviation (examples B.1 and B.2.) The standard deviation calculated from the results of different persons is more informative than that calculated from duplicate counts by one person. (One person can usually duplicate the count, however wrong, with considerable precision.) Systematic differences between persons should also be detected and evaluated.

The observations on counting uncertainty and time-sensitivity will give the first indication of potential problems with wide use of the method.

10.2.3 Other robustness features

If there are doubts about the effects of incubation temperature, moisture or gas atmosphere, they should be studied by special tests involving incubation of parallel subsamples under the alternative conditions.

It is recommended that decision in these cases be based on statistical tests rather than mere superficial impressions. Experimental designs based on factorial plans of the analysis of variance are usually the most appropriate.

Factors not specifically tested are considered ineffective or under control.

10.2.4 Upper working limit

Having established the "environmental" conditions and target characteristics, the next step is to explore the quantitative limitations of the detector in question.

This can be done with a single type of experiment: a finely graded series of dilutions or volumes, with parallels [21, 25]. The excercise provides the data for assessing the upper limit on the basis of proportionality and repeatability (annex C).

10.2.5 Precision

The laboratory responsible for introducing a new method should also be responsible for providing the initial values of its precision estimates. These should include estimates of counting uncertainty and basic repeatability and reproducibility estimates of different repetitions (parallel plates, parallel dilution series, matrices). Other laboratories need this information for their secondary validation of the method and subsequently for establishing the systems of analytical quality control.

If sufficient grounds are not found for calculating a Type B estimate from assumed or known statistical distributions and other information, split-sample or parallel-sample experiments shall be made to obtain Type A estimates of precision.

Collaborative method performance studies [18] involving several laboratories are inappropriate for the validation of new methods because they presume every participant to have solid experience with the method. Only later will this become possible.

10.2.6 Trueness and relative recovery

Absolute recovery (trueness) of the microbiological analyte is unmeasurable, which is a problem in the primary validation of a new method.

True recovery can be approximated with tests on pure cultures or spiked sterilized samples using a non-selective method as reference. Some certified reference materials are also available for the purpose.

In order to observe the relative recovery, natural samples are studied with the candidate method and a reference method in parallel. Low recovery compared with the recovery on a standard method is an obvious reason to doubt the method. To be valid the comparison must be based on confirmed colony counts.

10.2.7 Specifications

The numerical performance characteristics observed shall be written in a section of specifications in the method protocol in accordance with 7 c).

10.3 Secondary validation

A number of natural samples shall be obtained. Split samples or replicate dilution series are studied with parallel plating. Duplicate counting is practiced in order to verify expected counting performance.

A number of presumptive positives (at least 100) shall be isolated and verified. Categorical performance characteristics are calculated and compared with the specified values, if available.

11 Designs for determining specifications

11.1 A general model for basic quantitative specifications

An experimental design based on a finely graded series of dilutions or volumes with replication of plating and counting provides the data for determining the upper working limit and some other quantitative specifications (Example B.1, annex C).

The appropriate design is a reduced version [27] of a plan used as an analyst performance test [33].

A carefully mixed liquid sample or a suspension of a solid material is prediluted to a density giving an expected colony number per plate that somewhat exceeds the assumed upper limit of the detector performance. A series of six or seven further dilutions with dilution steps 1:2 is continued from the starting suspension. Three parallel plates are seeded from each dilution.

Plates from dilutions averaging more than 20 colonies per plate are read in a randomly coded order by one person.

If all or a randomly selected subset of the plates were read by a second person, the results would provide data on counting uncertainty as well. With many methods, the typical appearance of target colonies might change during the time required to carry out duplicate counting of an extensive series of plates. Counting agreement/disagreement should then be tested separately.

NOTE The geometric series based on dilution steps 1:2 is rather coarse. The colony numbers at the beginning decrease in large steps. Methods with a low upper working limit cannot be efficiently evaluated. A dilution ratio smaller than 1:2 is recommended in such cases.

Membrane filtration methods allow easy variation of sample volume, permitting a more appropriate gradation. In such instances an arthmetic series with volume ratios 1:2:3:4:5:6:7 is practicable. It has the additional advantage that all test portions are derived from a single suspension.

The data are analysed for proportionality, overdispersion of parallels and counting uncertainty, assuming perfect randomness at every step as the basis of evaluation.

The unit design (Example B.1, annex C) involves only one sample. A similar plan should be repeated with many samples (minimum ten) to be able to generalize the results.

The statistical calculations are based on procedures detailed in annex A, with practical examples B.3 and B.4 in annex B.

11.2 Precision of the entire analytical procedure

The plan described in 11.1 does not address the precision of the entire analytical procedure. To determine the precision of the determination, tests with split samples or replicate dilution series are required (Example B.2, annex C).

Natural samples with medium or high numbers of the analyte should be obtained. After a standard mixing procedure, replicate determinations are made. Parallel plating is recommended but not absolutely necessary.

At least thirty samples encompassing the entire scope are required for adequate coverage.

The number of parallel determinations (dilution series) per sample shall be at least two. Five or six parallels give more trustworthy results. Although recommended, the number of parallels need not be the same for every sample

The results are used for calculating the overdispersion constant (6.2.3) in the manner illustrated for parallel plates in example B.7.

11.3 Categorical characteristics

Characteristics related to selectivity, sensitivity, specificity etc. (9.2) can be studied in connection with the tests described in 11.2. unless working time becomes a limiting factor.

Plates with low, i.e. technically and biologically reliable, numbers of colonies are selected. All colonies of target and non-target types are counted.

In primary validation, colonies of both types are randomly isolated for verification. The number should be "not small", which means at least 20 or 30 of both types per sample, if available. The two colony types need not necessarily be from the same dilutions if selectivity does not permit it. When selectivity is high (target colonies more than 90 %) it may be unnecessary to isolate presumptive negative colonies.

In secondary validation, only presumptive positive colonies need be isolated and verified.

11.4 Unplanned data

Results from various parallel observations collected over time without special design can be used as additional information to support or amend upper working limit or overdispersion specifications.

Annex A

Statistical procedures and computer programs

A.1 Uncertainty of counting

Uncertainty of counting is estimated by reading the same plates more than once. From these results, the standard deviation or relative standard deviation can be computed.

NOTE As three-letter acronyms such as RSD may be confusing in mathematical formulae, the letter u is instead consistently used as the symbol for relative standard deviation.

Assume a plate with the (unknown) number x of characteristic colonies, and denote with x_1, x_2,..., x_n the numbers observed in counting of the plate by the same person repeatedly or by different persons.

Counting uncertainty u_Z is expressed as the relative standard deviation:

$$u_Z = \frac{s(x)}{\bar{x}}$$

where

$$s(x) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

In case of duplicate counting, the same can be calculated from

$$u_Z = \frac{\sqrt{2}\,|x_1 - x_2|}{x_1 + x_2}$$

A large number of cases (plates) with at least 20 colonies each should be randomly picked over time, not selecting unusual ones. The quadratic mean is taken as the estimated average relative counting uncertainty.

$$u_Z = \sqrt{\frac{u_{Z1}^2 + u_{Z2}^2 ... + ... u_{Zn}^2}{n}}$$

For reliable estimation, n should be at least 30.

Depending on the persons participating, the following terminolgy applies:

— **individual (or personal) repeatability:** a summary of the results of one person reading the same plates twice (or more);

— **within-laboratory repeatability:** a weighted (quadratic) mean value of the individual repeatabilities of the laboratory personnel; a summary of the duplicate or replicate counting results of all persons;

— **within-laboratory reproducibility:** the (quadratic) mean of results of several persons from the same laboratory reading the same plates;

— **between-laboratories reproducibility:** the (quadratic) mean of results of persons from different laboratories reading the same plates.

See Examples B.1, B.2 and B.3.

A.2 General test of proportionality (linearity)

Let n colony counts c_1, c_2,..., c_n be obtained from the study of the same test suspension in volumes or dilutions that are related as the numbers R_1, R_2,..., R_n

The log-likelihood ratio estimate of the proportionality (linearity) of the counts can be calculated from

$$G^2_{n-1} = 2\left[c_1 \ln\frac{c_1}{R_1} + c_2 \ln\frac{c_2}{R_2} + ... + c_n \ln\frac{c_n}{R_n} - \left(\sum c\right) \times \ln\left(\frac{\sum c}{\sum R}\right)\right]$$

A BASIC program for calculating G^2 is attached.

A guide value can be obtained by referring to tables of χ^2 with $n-1$ degrees of freedom. Values exceeding the tabulated value indicate departure from proportionality at the chosen probability level.

NOTE This formula can be used for testing the agreement of parallel plates as well by giving the same number (e.g. $R_1 = R_2 = ... = R_n = 1$) for each relative volume. The result is usually almost the same as with the Poisson index of dispersion given in A.3.

A BASIC program for calculating the general log-likelihood ratio index G^2 according to A.2. See Example B.5.

```
10    PRINT "LIKELIHOOD RATIO INDEX G^2"
20    INPUT "NUMBER OF TERMS, n= ";N
30    C=0: R=0: S=0: T=0: D=0
40    FOR I=1 TO N
50    PRINT "I= ";I
60    INPUT "COLONY COUNT= ";C
70    INPUT "RELATIVE VOLUME= ";R
80    IF C=0 THEN W=0: GO TO 100
90    W=C*LOG(C/R)
100   S=S+W
110   T=T+R
120   D=D+C
130   NEXT I
140   Y=2*(S-D*LOG(D/T))
```

```
150   Y=(INT(1000*Y+0.5))/1000
160   PRINT
170   PRINT "INDEX G^2= ";Y
180   PRINT
190   PRINT
200   INPUT "ANOTHER SET? (Y/N)";A$
210   IF A$="Y" OR A$="y" GOTO 20 ELSE 220
220   END
```

NOTE In this BASIC version LOG means the natural logarithm. In other versions the symbol LN may be required in rows numbered 90 and 140.

A.3 Poisson Index of Dispersion

Parallel colony counts c_1, c_2, ..., c_n from equal portions of a fully mixed suspension can be tested for randomness by calculating the Poisson Index of Dispersion:

$$X^2_{n-1} = \frac{n\sum c_i^2 - \left(\sum c_i\right)^2}{\sum c_i} = \frac{n\sum c_i^2}{\sum c_i} - \sum c_i$$

NOTE 1 Symbols D^2, χ^2, and T_1 are frequently also used for X^2.

NOTE 2 The general G^2 formula in A.2 can be applied instead.

Excess scatter (overdispersion) can be detected by referring to a table of χ^2 distribution with n – 1 degrees of freedom.

An isolated X^2 value is of little importance, especially if the number of parallel plates (n) is small. It is recommended that a large number (at least 100) of sets of parallel plates be studied over a lengthy period. The samples should represent different sources and seasons.

X^2 values and their degrees of freedom are additive. The general agreement of several sets of parallels can be tested by comparing the sum of the X^2 values with the theoretical χ^2 value with degrees of freedom corresponding to the sum of the degrees of freedom. The test is extremely powerful.

When about 100 individual X^2 values are available, they can also be grouped in frequency classes, with class boundaries read from the appropriate χ^2 table. The observed frequencies can be compared with the frequencies expected according to the theoretical distribution. Comparing the frequency distributions of two methods in this way is more informative than the mere sums. (See Example B.6.)

A BASIC program for calculating the Poisson Index of Dispersion X^2 given in A.3. See also Example B.6.

```
10    PRINT "DISPERSION INDEX X^2"
20    INPUT "NUMBER OF PARALLELS, n= ";N
30    S1=0: S2=0
40    FOR I=1 TO N
50    PRINT "I= ";I
60    INPUT "COLONY COUNT= ";C
70    S1=S1+C
80    S2=S2+C^2
90    NEXT I
100   X = (N*S2-S1^2)/S1
110   X2=(INT(1000*X+0.5))/1000
120   PRINT
130   PRINT "INDEX X^2= ";X2
140   PRINT
150   INPUT "ANOTHER SET? (Y/N)";A$
160   IF A$="Y" OR A$="y" GOTO 20 ELSE 170
170   END
```

Annex B

Numerical examples

B.1 General

Examples B.1, B.2 and B.3 deal with repeatability and reproducibility of counting, i.e. the results of reading the same plates by one or more persons. Example B.4 is connected with the robustness of results with respect to incubation time. Example B.5 illustrates an analysis of linearity, Example B.6 shows ways to use the information contained in parallel plates and Example B.7 illustrates the calculation of the overdispersion constant from parallel test data.

The examples presented in B.1 to B.3 illustrate the calculations recommended whenever duplicate or replicate counting results are available. Such results are based on reading the same plates repeatedly under uniform conditions, i.e. within a time interval clearly shorter than the assumed tolerance allowed for the method. In practice this means a maximum interval of 1 h.

Depending on the persons participating, the same formal calculations (see A.1) yield estimates of different coverage.

The plates for repeated counting should be selected at random, ignoring plates with less than 20 colonies. Otherwise, the plates picked should represent the whole application range of the method.

For a reasonably reliable general estimate, at least 30 cases should be available.

The between-laboratories repeatablity is best studied by transporting people to common sessions rather than sending plates around.

NOTE In the easiest microbiological counting task, that of counting colonies of a pure culture with regular medium-sized colonies on membrane filters, the count can be expected to be repeatable with a standard deviation better than 2 % (RSD < 0,02). This applies almost equally when the counting is repeated by one person or by different persons from different laboratories, up to several hundred colonies per plate.

When the target colonies must be identified in a mixed population on the basis of colour, size, shape, texture or reactions with the growth medium, the task is more difficult. One person can often repeat the count quite accurately but the deviations between different persons become large. How large an uncertainty is tolerable in a valid method has not been decided. Relative standard deviation greater than 0,1 (five to ten times the RSD of pure-culture counting) is a certain sign of problems or difficulties.

B.2 Example B.1 — Personal repeatability of counting

B.2.1 General

In the following example of real data on total colony counts on non-selective media, the repeatability relative standard deviation was mostly larger than the "ideal" (RSD < 0,02) referred to above.

Persons A and B, working in the same laboratory independently, read different plates twice within a short time interval. The duplicate counts are denoted by x_1 and x_2.

B.2.2 Calculations

The mean (m) and standard deviation (s) of each pair are first computed. From them the values of RSD = s/m are obtained (penultimate column).

Plate	x_1	x_2	m	s	RSD	Person
1	129	122	125,5	4,95	0,0394	A
2	417	377	397,0	28,28	0,0712	A
3	73	80	76,5	4,95	0,0647	A
4	49	52	50,5	2,12	0,0420	A
5	86	81	87,5	3,54	0,0423	B
6	37	39	38,0	1,41	0,0372	B
7	112	115	113,5	2,12	0,0187	B
8	204	214	209,0	7,07	0,0338	B
9	66	71	68,5	3,54	0,0516	B
10	306	299	302,5	4,95	0,0164	B

The personal mean repeatability estimates are obtained as follows.

Person A:

$$\mathrm{RSD}_{\mathrm{A}} = \sqrt{\frac{0{,}039\,4^2 + 0{,}071\,2^2 + 0{,}064\,7^2 + 0{,}042\,0^2}{4}} = 0{,}056$$

Person B:

$$\mathrm{RSD}_{\mathrm{B}} = \sqrt{\frac{0{,}042\,3^2 + \ldots + 0{,}016\,4^2}{6}} = 0{,}036$$

To obtain the weighted pooled relative standard deviation of repeatability in the laboratory where A and B are the technicians, the quadratic mean of all ten values can be calculated in the same way:

$$\mathrm{RSD}_c = \sqrt{\frac{0{,}039\,4^2 + 0{,}071\,2^2 + \ldots + 0{,}016\,4^2}{10}} = 0{,}046$$

An unweighted pooled estimate might be more appropriate. It can be obtained by taking the quadratic mean of the personal means:

$$\mathrm{RSD}_c = \sqrt{\frac{\mathrm{RSD}_{\mathrm{A}}^2 + \mathrm{RSD}_{\mathrm{B}}^2}{2}} = \sqrt{\frac{0{,}056^2 + 0{,}036^2}{2}} = 0{,}047$$

B.3 Example B.2 — Pooled repeatability estimate: Analysis of variance

The mean values for one person or the laboratory as a whole can also be obtained by one-way analysis of variance. Colony counts x_1 and x_2 in Example B.1 are first converted to their ln-values and the between- and within-plate variances are computed.

Analysis of variance of all repeatability results is shown below:

	DF	SS	MS
Between-plates	9	10,481 7	1,164 6
Within-plates	10	0,019 0	0,001 9

where

DF = degrees of freedom;

SS = sum of squares;

MS = mean square (variance).

The within-plates MS represents the pooled weighted repeatability.

The relative standard deviation for repeatability is the square root of the within-plates mean square (because standard deviation in ln-scale roughly corresponds to relative standard deviation in arithmetic scale).

$$\mathrm{RSD}_c = \sqrt{0,001\,9} = 0,044$$

The value is very nearly the same as in Example B.1.

If the value is high (larger than 0,1) it may be worth returning to the table to examine the individual RSD values in search for reasons. One accidental large value may be responsible, or a trend on the mean colony count may be present. They are best illustrated graphically by plotting the RSD values against the colony count.

The personal or within-laboratory repeatability of counting is needed as a base value when estimating other robustness features such as time-sensitivity or reproducibility of counting. The ideal value of 0,02 is frequently too optimistic.

B.4 Example B.3 — Between-laboratories reproducibility of counting

Two persons (A1, A2) from laboratory A and three (B1, B2, B3) from laboratory B joined in a colony-counting session arranged by laboratory B. Standard agar plates were picked from routine determinations and were read by each participant. Plates with less than 30 colonies were omitted. Results of six plates are shown in the list below.

Plate	A1	A2	B1	B2	B3	Mean	RSD
1	33	26	33	34	33	31,8	0,102 9
2	160	156	166	176	174	166,4	0,052 0
3	142	128	142	146	139	139,4	0,049 1
4	78	97	81	81	83	84,0	0,089 1
5	89	94	81	94	92	90,0	0,060 3
6	38	44	38	42	40	40,4	0,064 5

Six plates are far too few for a reliable general estimate, but illustrate the computations.

The quadratic mean of the relative standard deviations is

$$RSD_c = \sqrt{\frac{0{,}102\,9^2 + 0{,}052\,0^2 + 0{,}049\,1^2 + 0{,}089\,1^2 + 0{,}060\,3^2 + 0{,}064\,5^2}{6}} = 0{,}072\,4$$

Compared with the reproducibility of counting within one laboratory (Examples B.1 and B.2) the value obtained in this collaborative experiment was almost twice that value. Analysis of variance may be employed in the search of systematic differences between persons or between laboratories.

B.5 Example B.4 — Robustness: Time-sensitivity of colony counts

Dependence of the colony count on incubation time can be easily tested by counting the same plates twice: at the two extreme incubation times permitted by the method.

In the total coliform analysis by the membrane filtration (MF) method according to ISO, the membranes should be incubated for 18 h to 24 h at a specified temperature.

To test the validity of the given limits, five water samples were tested by the MF method. Plates with "reliable" numbers of colonies were counted after 18 h and 24 h.

The results are listed below.

Sample	Count 18 h	Count 24 h	Change %
1	90	93	+ 3
2	44	88	+ 100
3	70	69	- 2
4	8	49	+ 613
5	29	69	+ 238

In only two samples (1 and 3), the difference between 18 h and 24 h counts fits within the normal repeatability of counting (CV = 3 % to 4 %) illustrated by Examples B.2 and B.3.

There is usually no need to test results of such experiments statistically. In principle, one clear exception is sufficient to reject the hypothesis of robustness. In this example, every one of samples 2, 4 and 5 indicated that the count changes too much within the assumed acceptable range.

The result of this example should be interpreted so that either the specifications for incubation time tolerance or the scope of the method are reconsidered.

B.6 Example B.5 — Analysis of a proportionality test: Upper limit of detector performance

B.6.1 General

A natural sample was prediluted to suitable level whereafter a dilution series of six successive steps of 1:2 was prepared according to the plan presented in annex C.

Three parallel plates were made from each dilution using the surface-spread technique. The colonies were counted after two days' incubation.

Dilution	Parallel counts			Sum S_i	Relative volume $V_{R,i}$	Ratio $S_i/V_{R,i}$
2^{-1}	121	204	162	487	32	15,22
2^{-2}	109	128	148	385	16	24,06
2^{-3}	111	114	97	322	8	40,25
2^{-4}	56	60	68	184	4	46,00
2^{-5}	36	29	24	89	2	44,50
2^{-6}	11	13	17	41	1	41,00
			Total:	**1508**	**63**	

B.6.2 General proportionality test

To test the general proportionality ("linearity") of colony counts and sample volume, it is sufficient to calculate the log-likelihood ratio index (G^2) for the agreement of the sums of parallel colony numbers with the respective relative volumes.

The counts should agree with the geometric series 32/16/8/4/2/1.

The test statistic has 6 – 1 = 5 degrees of freedom.

According to the formula (annex A, item A.2):

$$G_5^2 = \left[487 \ln(487/32) + 385 \ln(385/16) + 322 \ln(322/8) + 184 \ln(184/4) + \right.$$
$$\left. + 89 \ln(89/2) + 41 \ln(41/1) - 1508 \ln(1508/63) \right] = 292{,}526$$

The value of the index is compared with the χ^2 distribution with five degrees of freedom. The calculated value exceeds the theoretical value for 0,1 % (20,515), which means that the general linearity of the results is extremely poor. (The conclusion is obvious in this case even without calculation, and can be reached by simply looking at the sums in the table.) Obviously the ratio 2:1 between successive dilutions is not realized in the colony counts.

The concluson is, therefore, that the detector system was not linear in this sample in the colony count range from 41/3 = 14 to 483/3 = 161 colonies per plate. From the inspection of the sums, or even more clearly of the counts per relative volume (S_i/V_i), it can be concluded that the high colony numbers deviate the most from expectation. The detector has become disfunctional even below 160 colonies per plate.

With similar proportionality tests performed on different samples, a number of paired values can be obtained: highest mean count tested (161 in this case) and the dispersion index value observed (G^2 = 292,526 in this case). Plotting the index values on guidance charts helps determine the highest colony count where proportionality of the method is sufficient.

An example of 14 samples examined in this way is shown in Figure B.1. An arbitrary guideline value of 15,09 (1 % probability) was chosen and is shown by a horizontal line. Points above the line belong to series with poor proportionality. The horizontal axis (c_{max}) gives the mean count per plate of the lowest dilution included in the proportionality test. The horizontal and slanted lines intersect at the mean colony number where the probability of "adequate" proportionality becomes less than 1 %.

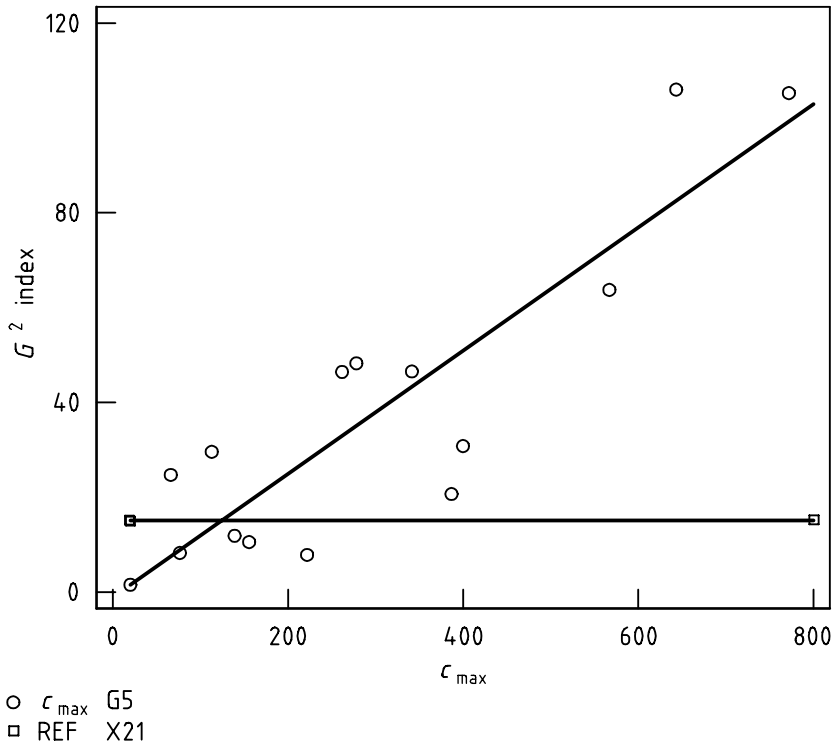


Figure B.1 — Linearity of series of counts from six successive binary dilutions of 14 samples

Linearity measured in terms of $G^2 \cdot c_{max}$ = mean colony count per plate of the dilution with highest colony numbers.

The conclusion is not always as obvious as in this example. A more detailed analysis of the data may be needed in order to explain the lack of proportionality.

NOTE 1 The use of the above reasoning presupposes that all series included in the proportionality study have the same number of countable dilutions (degrees of freedom). If it should happen that all series do not have the same number of countable dilutions, another proportionality measure $I = G^2/(n-1)$ should be used. The guide value becomes somewhat more vague in this instance because $I = \chi^2/(\upsilon)$ is not a constant but depends on the degrees of freedom (υ).

NOTE 2 $I = \chi^2/(\upsilon)$ used to be called the Lexis ratio, in honour of the German economist Lexis. It was frequently denoted by the symbol L.

To test in greater detail where the loss of linearity takes place, the total index G_5^2 = 292,526 can be subdivided in orthogonal comparisons by working up or down the table and forming successively the contrasts for each individual sum, with all the sums below it taken together.

B.7 Example B.6 — Parallel plate data

B.7.1 General

Data on parallel plate counts are easily collected. They are useful for method validation because the technical difference between parallel plates consists of almost nothing other than volume errors in pipetting. Unless something is radically wrong with the pipetting technique, the volume uncertainty practically vanishes among the random particle scatter (6.1).

Statistical agreement (randomness) of a set of parallel plates can be measured by calculating the Poisson index of dispersion according to A.3 or the log-likelihood ratio index according to A.2.

If substantial disagreement with the Poisson randomness is observed, the cause is almost certainly something other than inaccuracy of pipetting. In this way the analysis of parallel plate data can serve the purposes of method validation.

An individual index value can be used for assessing the statistical reliability of that particular parallel plate set. An excessively high index value, compared with the χ^2 distribution, may indicate that the mean of that set is not reliable. In this way the Poisson index of dispersion is an important AQC tool.

More importantly, index values collected from a large number of different samples can be used for guidance on aspects of method performance. Data on two methods gives an opportunity to compare their relative reliability, and secondary validation will benefit from a comparison with specifications developed during primary validation.

B.7.2 Frequency distributions of Poisson indices

The following table shows the first few rows of data on parallel counts (B1, B2) of sixty samples. They represented a wide range of particle concentrations. The Poisson index of dispersion (X^2) was calculated for each parallel set according to A.3.

Suspension	B1	B2	X^2
1	256	302	3,792
2	228	146	17,979
3	89	108	1,832
4	27	29	0,071
5	143	129	0,721
etc.			

For instance, for the first pair of results the X^2 value is obtained from the calculation:

$$X^2 = \frac{2\left(256^2 + 302^2\right)}{256 + 302} - \left(256 + 302\right) = 3{,}792$$

The index has a theoretical statistical distribution. The observed distribution of indices can be compared with the theoretical expectation by noting the frequency of index values of different size. The class boundaries are obtained from the percentage points of the χ^2 distribution. To do that, the percentage points of the χ^2 distribution with the appropriate degrees of freedom (one less than the number of parallels) must be consulted. In this case the table for one degree of freedom is needed.

Percentage points P of the χ^2 distribution for one degree of freedom:

P	0,95	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05
χ^2	0,004	0,016	0,064	0,148	0,256	0,455	0,722	1,07	1,62	2,71	3,84

The simplest way to use the table is to construct class boundaries for an even distribution of expected frequencies. For instance, of all index values, 20 % are expected to fall by chance in each of the five classes with the following boundaries:

Class 1	0 to 0,064
Class 2	0,064 to 0,256
Class 3	0,256 to 0,722
Class 4	0,722 to 1,62
Class 5	> 1,62

The expected frequency in each class depends on the number of cases studied. With 60 cases (the example below), the expectation would be that 12 observations fall in each class.

Other than even distributions can be constructed in a similar manner. A different construction is shown in the example below. The data were divided in six classes, with 5 %, 15 %, 30 %, 30 %, 15 % and 5 % cases expected in the respective classes.

The data shown in the example case above are for one particular method (Method A). Another method (Method B) was tested simultaneously on the same suspensions in exactly the same way. The dispersion indices of the parallel plates with that method were also calculated and collected in the frequency classes. Results are presented in Table B.1.

Table B.1 — Comparison of the observed frequencies of dispersion indices of two methods with the theoretical expectation

Frequency class	Upper class boundary	Expected frequency	Observed frequency	
			Method A	Method B
1	0,004	3	1	2
2	0,064	9	3	7
3	0,455	18	10	10
4	1,64	18	15	22
5	3,84	9	10	10
6	> 3,84	3	21	9
	Total	**60**	**60**	**60**

Obviously the parallel determinations of method B are closer to the theoretical expectations. The fit could be tested by Pearson's classical goodness-of-fit test in case agreement with Poisson distribution has been indicated in the specifications. If methods A and B are otherwise equivalent, the observation clearly favours the choice of method B.

A more detailed inspection of the index values might reveal reasons for the excessive amount of high values with method A. If found correlated with colony counts, it could be used as additional information about the upper working limit of the method.

B.8 Example B.7 — Computation of the overdispersion component u from parallel count data

Whenever parallel observations in the form of untransformed count data are available, the principle described in 6.2.3 [equation (12)] can be used for estimating the additional (overdispersion) component of the analytical procedure.

The following example is based on a multicentre method performance test. More than 50 laboratories participated. Results of twelve laboratories were picked to illustrate the calculations.

Each laboratory studied a sample of their own. It was only agreed beforehand that the sample must represent the same type of material (municipal waste water). All used the same analytical procedure except that each laboratory had complete freedom as to how they mixed and diluted the sample they had taken.

As a consequence, the results of different laboratories had widely different mean colony counts in the dilutions they had available for counting.

The experimental design consisted of four dilution series made from the homogenized sample suspension. One plate per dilution was made. The laboratories were instructed to choose the same dilution in all four series for counting.

Table B.2 — Results of different laboratories

Laboratory	Parallel results				c	Variance	VTM
	C1	C2	C3	C4			
1	198	233	218	254	225,8	560,2	2,48
2	155	145	150	131	145,3	106,9	0,75
3	58	53	64	66	60,3	34,9	0,58
4	37	42	38	31	37,0	20,7	0,56
5	124	106	92	117	109,8	194,9	1,78
6	28	17	11	20	19,0	50,0	2,63
7	167	238	213	206	206,0	864,7	4,20
8	10	12	13	8	10,8	4,9	0,46
9	66	84	94	71	78,8	160,9	2,04
10	8	13	7	5	8,3	11,6	1,40
11	204	186	225	216	207,8	284,2	1,37
12	162	141	166	199	167,0	575,3	3,45

c = mean count per plate.

VTM = variance-to-mean ratio ("Lexis ratio").

According to the principle given in 6.2.3, the regression line of VTM (= Y) on c is expected to have the form:

$$Y = 1 + u^2 c$$

The slope of the line thus gives an estimate of the squared overdispersion constant.

The means and variances were based on a small number (four) of parallels. Considerable random scatter is to be expected in their ratio estimates. This fact is clearly evident when the results of this experiment are plotted (Figure B.2).

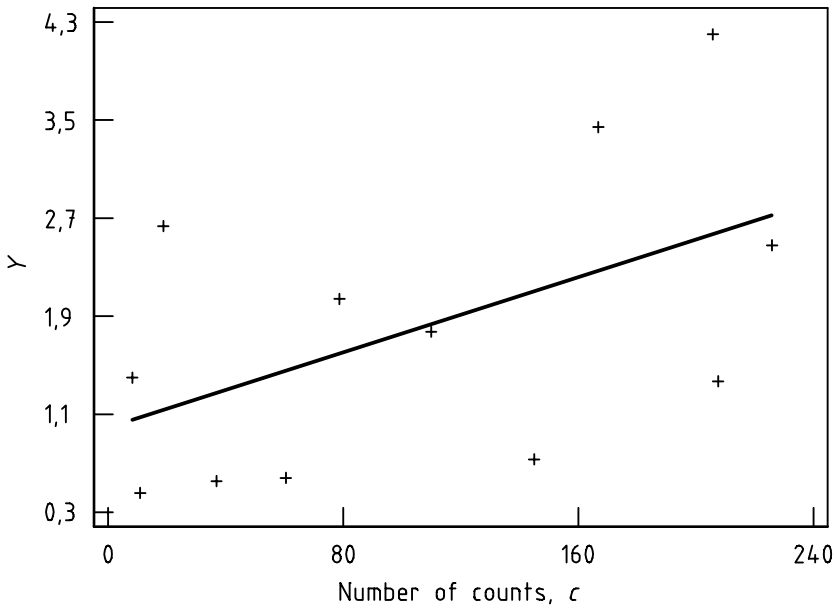


Figure B.2 — Dependence of the variance-to-mean ratio on the mean colony count in an experiment with wastewater samples

Each point represents one sample studied by a different laboratory. The means and variances were based on untransformed colony counts in four parallel determinations of a test suspension.

Because of the large scatter, considerable doubt remains whether the regression line accurately depicts the true relationship. More data would be needed to estimate the slope with a high degree of confidence. The calculation can nevertheless be continued, to illustrate the method.

The least squares regression line fitted to the data points has the equation:

$$Y = 0{,}99 + 0{,}007\,66c$$

The constant (0,99) happens to be very near the expected value of 1,0. The slope gives the squared overdispersion value $u^2 = 0{,}007\,66$. The excess variation as relative standard deviation is therefore $u = 0{,}088$.

The slope is not statistically significant. It is not convincingly demonstrated that there are sources of variation in addition to the Poisson distribution. It is not advisable to proceed much further in the interpretation. The result is however interesting enough that one might conclude it worth the effort to gather more similar data to be able to estimate the slope with higher confidence.

NOTE It depends on the experimental plan what components of uncertainty the overdispersion constant contains. The main factors that might have caused additional variation in this particular experimental plan were:

— possible lack of perfect mixing of the samples;

— volumetric random errors in dilutions;

— personal counting uncertainty.

The value of the constant 0,088 means that the added component is only about ± 9 %. The heterogeneity of the sample, cumulative volumetric uncertainties and the personal counting repeatability combined may be interpreted to have remained within acceptable limits.

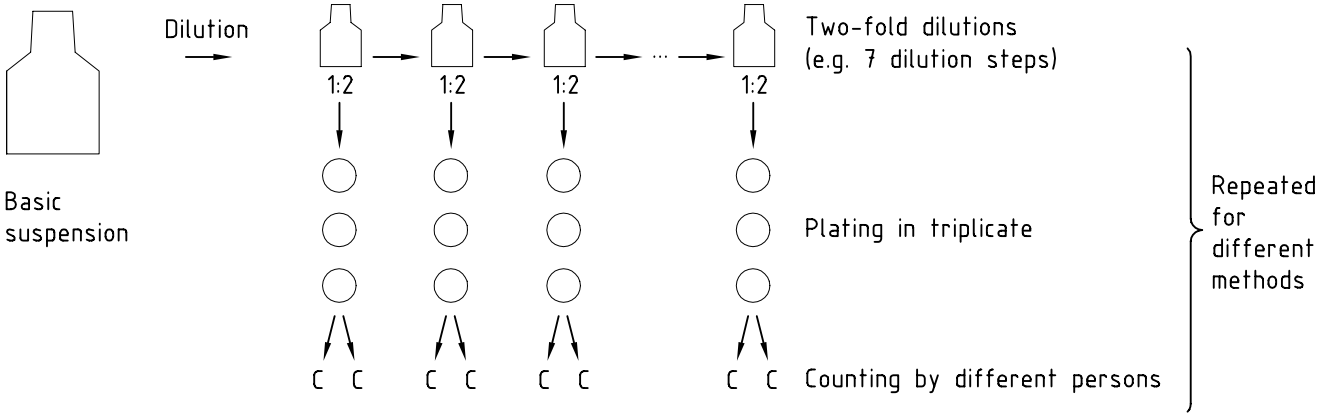
In this plan, where every laboratory used a sample of their own, nothing can be learned about the relative proficiency of the different laboratories. All differences in relative recovery, interpretation of colony appearance and functioning of nutrient media have disappeared. To include these factors, all laboratories should have studied an identical sample and followed identical procedures.

If the parallel results had been parallel counts from the same final dilution flask, then the overdispersion factor would not have included material inhomogeneity and dilution uncertainty. The only remaining additional sources would have been pipetting and counting.

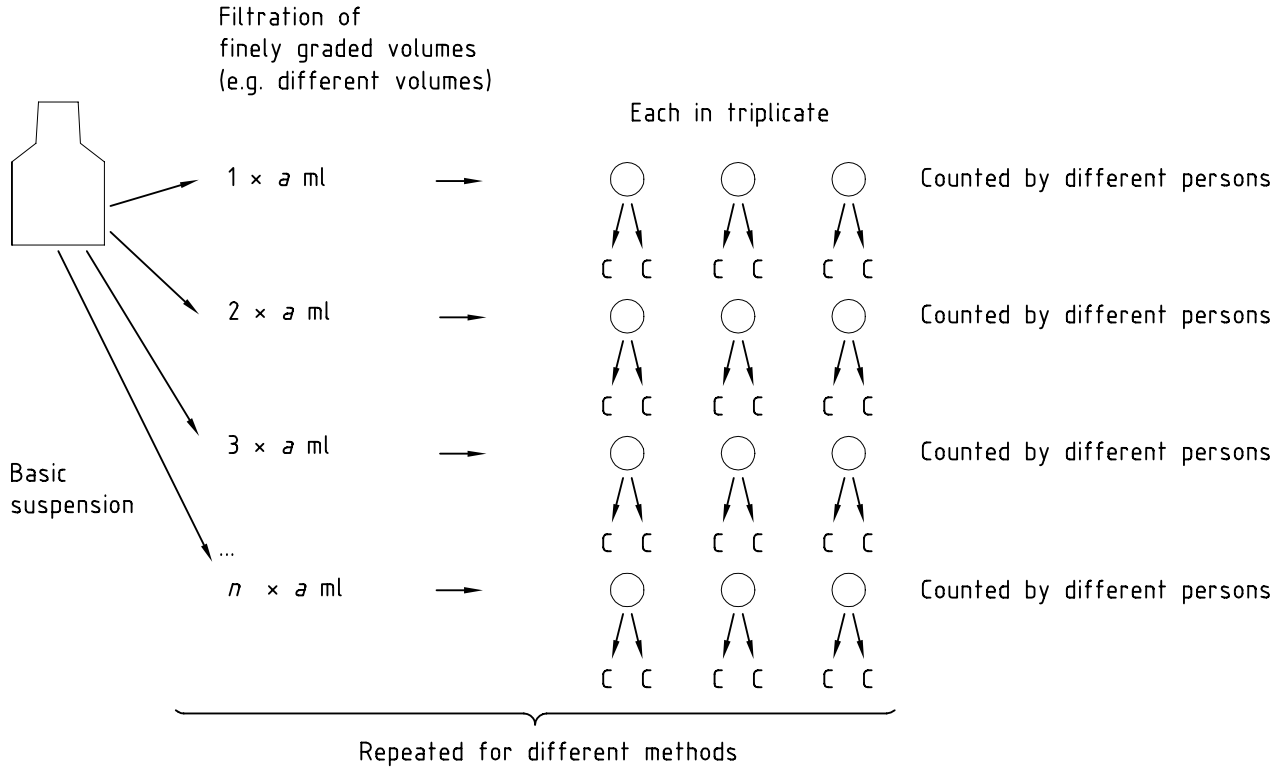
Annex C

Example of a validation experiment

C.1 Plate count and spread plate (repeated for different samples)



C.2 Membrane filtration (repeated for different samples)



NOTE Any data on replicate counting or parallel plates available can be used for study of specific details of validation. Sensitivity and selectivity should be proven before use of this design.

Bibliography

[1] FAO Manual of Food Quality Control 12. *Quality assurance in the food control microbiological laboratory.* (FAO Food and Nutrition Paper 14/12. Rome, 1991. 154 pp.)

[2] Report on Public Health and Medical Subjects No. 71. *Methods for the Examination of Waters and Associated Materials. The Microbiology of Water. Part 1 — Drinking Water.* HMSO, 1994.

[3] ISO 3534-1:1993, *Statistics — Vocabulary and symbols — Part 1: Probability and general statistical terms,* and ISO 5725 (all parts), *Accuracy (trueness and precision) of measurement methods and results.*

[4] ISO 9998:1991, *Water quality — Practices for evaluating and controlling microbiological colony count media used in water quality tests.*

[5] *Quantifying uncertainty in analytical measurement.* 1995. Eurachem. ISBN 0-948926-08-2

[6] *Guide to the expression of uncertainty in measurement.* (GUM), 1995. International Organization for Standardization, Geneva. ISBN 92-67-10188-9

[7] *Handbook for Microbiological Laboratories. Introduction to Internal Quality Control of Analytical Work.* Nordic Committee on Food Analysis (NMKL), Report No. 5, 1989.

[8] *Standard Methods for the Examination of Water and Wastewater*, 18th ed. 1992. APHA, AWWA, WEF.

[9] ISO 7218:1996, *Microbiology of food and animal feeding stuffs — General rules for microbiological.*

[10] prEN 275-061:1997 *E. Microbiology of food and animal feeding stuffs — Protocol for the validation of alternative methods.*

[11] ANSCOMBE, F.J. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37,** 1950, pp. 358-382.

[12] COCHRAN, W.G. Estimation of bacterial densities by means of the "Most Probable Number". *Biometrics*, **6**, 1950, pp. 39-52.

[13] COCHRAN, W.G. *Sampling Techniques*, 3rd edn. John Wiley & Sons, New York, 1977.

[14] EL-SHAARAWI, A.H., ESTERBY, S.R. and Dutka, B.J. Bacterial density in water determined by Poisson or negative binomial distributions. *Appl. Environ. Microbiol.*, **41**, 1981, pp. 107-116.

[15] FISHER, R.A. The negative binomial distribution. *Ann. Eugenics*, **11**, 1941, pp. 182-187.

[16] FISHER, R.A., THORNTON, G.G. and MACKENZIE, W.A. The accuracy of the plating method of estimating the density of bacterial populations. *Ann. Appl. Biol.*, **9**, 1922, pp. 325-359.

[17] HAVELAAR, A.H., HEISTERKAMP, S.H., HOEKSTRA, J.A. and MOOIJMAN, K.A. Performance characteristics of methods for the bacteriological examination of water. *Water Sci. Technol.*, **27**, No. 3-4, 1993, pp. 1-13.

[18] HORWITZ, W. 1988. Protocol for the design, conduct and interpretation of collaborative studies. *Pure & Appl. Chem.*, **60**, pp. 855-864.

[19] HURLEY, M.A. and ROSCOE, M.E. Automated statistical analysis of microbiological enumeration by dilution series. *J. Appl. Bacter.*, **55**, 1983, pp. 159-164.

[20] JARVIS, B. Statistical aspects of the microbiological analysis of foods. *Progr. Ind. Microbiol.*, **21**, 1989, p. 179.

[21] LIGHTFOOT, N.F. and MAIER, E.A. (eds.). *Microbiological analysis of food and water. Guidelines for quality assurance.* Elsevier, Amsterdam, 1998, 266 p.

[22] LIGHTFOOT, N.F., TILLETT, H.E., BOYD, P. and EATON, S. Duplicate split samples for internal quality control in routine water microbiology. *Lett. Appl. Microbiol.*, **19,** 1994, pp. 321-324.

[23] MASSART, D.L., VANDEGINSTE, B.G.M., DEMING, S.N., MICHOTTE, Y. and KAUFMAN, L. *Chemomertics: a textbook. Data handling in science and technology.* Vol. 2. Elsevier, Amsterdam, 1988, 488 p.

[24] MCCLURE, F.D. Design and analysis of qualitative collaborative studies: minimum collaborative program. *J. Assoc. Off. Anal. Chem.*, **73**, 1990, pp. 953-960.

[25] MOSSEL, D.A.A., BONANTS-VAN LAARHOVEN, T.M.G., LIGTENBERG-MERKUS, A.M.T. and WERDLER, M.E.B. Quality assurance of selective culture media for bacteria, moulds and yeasts: an attempt at standardization at the international level. *J. Appl. Bacteriol.*, **54**, 1983, pp. 313-327.

[26] NIEMELÄ, S.I. Quantitative estimation of bacterial colonies on membrane filters. *Ann. Acad. Sci. Fenn., Ser. A. IV Biologica*, 1965.

[27] NIEMELÄ, S.I. A semi-empirical precision control criterion for duplicate microbial colony counts. *Lett. Appl. Microbiol.*, **22**, 1996, pp. 315-319.

[28] NIEMELÄ, S.I. *Performance characteristics of microbiological water analytical methods. European Training Courses on Water Quality Measurements. Course A: Monitoring and Measurements of Lake Recipients.* Helsinki, Finland 25-29 August 1997.

[29] STUDENT. On the error of counting with a haemacytometer. *Biometrika*, **5**, 1907, pp. 351-360.

[30] TILLETT, H.E., LIGHTFOOT, N.F. and EATON, S. External quality assessment in water microbiology: statistical analysis of performance. *J. Appl. Bacteriol.*, **74**, 1993, pp. 497-502.

[31] TILLET, H. E. and LIGHTFOOT, N.F. Quality control in environmental microbiology compared with chemistry: What is homogeneous and what is random? *Water Sci. Technol.*, **13**, 1995, pp. 471-477.

[32] TOMASIEWICZ, D.M., HOTCHKISS, D.K., REINBOLD, G.W., READ, R.B. Jr. and HARTMAN, P.A. The most suitable number of colonies on plates for counting. *J. Food Protect.*, **43**, 1980, pp. 282-286.

[33] WEISS, H. and DAHMS, S. Statistically based analyst performance assessment for microbiological analysis. Chapter 37 in: *Analytical Quality Assurance and Good Laboratory Practice in Dairy Laboratories*. International Dairy Federation Special Issue No. 9302, 1993. ISBN 92 9098 010 2.

[34] YOUDEN, W.J. and STEINER, E.H. *Statistical Manual of the AOAC.* Association of Official Analytical Chemists. Arlington, VA, USA, 1975, ISBN 0-935584-15-3.

ICS 07.100.20

Price based on 47 pages