FSIS Guidance for Test Kit Manufacturers, Laboratories: Evaluating the Performance of Pathogen Test Kit Methods

Table of Contents

I. Introduction

FSIS-regulated establishments rely on results from pathogen testing programs to comply with regulatory requirements and to support decisions made in their HACCP systems to ensure the production of safe unadulterated products. FSIS does not maintain a list of acceptable methods to be used in these testing programs. However, the Agency's overall expectation is that any test used by an establishment is appropriate for its intended use, that the test performance is comparable to the FSIS method (if applicable), and that the laboratory performing the test did not introduce modifications that could compromise test's performance.

FSIS believes that a robust validation study must be performed on any method used by establishments to detect microbiological hazards in FSIS-regulated foods. A validation study is an experimental process to measure performance characteristics of a particular test, with the goal of determining whether the test is equivalent to the reference test. "Equivalent" is defined as the designated relationship between two tests indicating that, for the intended conditions of use, the performance characteristics are statistically indistinguishable.

This guidance document (section II) provides an example of how to design a robust pathogen method validation study that may be used to demonstrate equivalence. The performance characteristics addressed in this guidance are described in Box 1. The guidance provided in this document primarily focuses on measuring relative recovery and sensitivity (false negative rate). Measurement of specificity (false positive rate), inclusivity, exclusivity, repeatability, reproducibility, and ruggedness should be performed through the direction of an independent organization, or by following guidance provided by the AOAC International Official Methods of Analysis Program[2]. The FSIS guidance should be useful to organizations that design or conduct validation studies for foodborne pathogen testing methods. These organizations include test kit manufacturers, laboratories, and independent validation organizations.

The FSIS guidance is not intended to conflict with or supplant existing guidance from independent organizations (AOAC and ISO).

The FSIS guidance could be used to evaluate the performance of a candidate alternative method for *Escherichia coli* O157:H7 or non-O157 Shiga toxin producing *E. coli* (non-O157 STEC).

NOTE: The use of validation in this document is not intended to have any application to the implementation of 9 CFR 417.4(a)(1) on initial validation of HACCP plans. This document deals exclusively with the evaluation of pathogen test kit methods.

Box 1. Performance Characteristics Used to Evaluate Pathogen Test Kit Methods

Relative Recovery measures the proportion of true positive samples recovered from the new test compared to the reference test when similarly inoculated.

Sensitivity/False Negative Rate measures the probability that a test will correctly detect a true positive sample. A false negative (FN) result occurs when a test does not correctly detect a true positive sample, so 1 minus the sensitivity equals the FN rate.

Specificity[1]/False Positive Rate measures the probability that a test will correctly detect a true negative sample. A false positive (FP) result occurs when a test does not correctly detect a true negative sample, so 1 minus the specificity equals the FP rate.

Inclusivity measures the ability of a test to detect a wide variety of strains representing the target pathogen.

Exclusivity measures the ability of a test to resist interference by cross-reactivity with non-target organisms likely to be found in the tested food.

Reproducibility is a measure of test performance in different laboratories with different equipment and personnel.

Repeatability is a measure of test performance in the same laboratory with the same equipment and personnel.

Ruggedness testing is performed to determine if small changes to the procedure or environmental factors influences test performance.

Validation studies are designed to evaluate the performance of a new test (referred in this document as the alternative method, or **A**) against a reference method (referred to as **R**) that provides a definitive result. The typical study design can not be applied to methods which do not have an available authoritative **R**. Additionally, the study design described in this document (section II) would not determine if the performance of **A** exceeded **R**.

Validation studies performed through the Association of Analytical Communities (AOAC) or other recognized independent organizations that perform or organize validation studies on behalf of test developers, follow the traditional design and rely on culture based reference methods. FSIS believes that any method used to detect foodborne pathogens in meat, poultry, and egg products should be as sensitive as the FSIS method. Other recognized cultural methods, fit for the purpose of detecting low levels of stressed cells in food, also may be an appropriate reference method. Alternative methods should be re-validated when significant changes affecting performance are introduced to the reference method. Re-validation should be performed within one year of the introduced changes.

[1] Also referred to as Selectivity.

From a food safety perspective, methods to detect foodborne pathogens should be validated using a robust study design with special attention to sensitivity (false negative rate) and inclusivity to limit or prevent false negative results. From an economic perspective, methods should be validated with special attention to specificity and exclusivity to prevent or limit false positive results and to reduce the time to obtain results, thus allowing product disposition to be rapidly determined. Sensitivity, specificity and timeliness are related, so an increase in one parameter may lead to a reduction of the other.

Robust validation methodology implies that **A** should be evaluated under the most challenging conditions to provide confidence that the method likely will perform well under most situations. In practical terms, a robust validation methodology should address the following parameters:

1. The inoculum level should be low enough to achieve fractional recovery of positive results by **R**. In FSIS' experience, pathogens subjected to zero tolerance testing in meat, poultry, and egg products often are found at low levels, close to one viable organism per analytical unit. Because it is practically impossible to place a single organism in a testable unit of food, the best approach taken in validation studies is to inoculate foods at low levels so that a fraction of the analyses (defined as 20-80% of the inoculated samples analyzed by **R**) are confirmed positive for the target pathogen. "Fractional recovery" is a well-established concept used by AOAC and other organizations performing validations, and was recognized as a preferred method for defining test performance by the Presidential Task Force for Best Practices in Microbiology[2,3].
2. The study should evaluate the ability of the test to detect potentially stressed or injured cells. Foods prepared for commercial distribution often are exposed to conditions injurious to bacterial contaminants. Foods often are processed at reduced temperatures to prevent pathogen growth and avoid spoilage. In other situations, food properties are modified by the application of antimicrobial agents such as organic acids, salt, curing agents, or other preservatives or by the modification of pH and water activity. In addition, the presence and level of resident microflora in the sample, (which is related to the age and handling of the product sample) could interfere with target pathogen growth. Any of these treatments may negatively affect the growth properties of the target pathogen, by extending lag phase or exponential growth rate. Injured cells would be more difficult to detect, but could retain their ability to cause illness.
3. The study should evaluate the ability of the test to detect target organisms in the products likely to be tested; however, foods that present a challenge to the test's performance should be evaluated, even if they are not as likely to be tested.
4. The study should evaluate a target strain with limited growth potential in the product to be tested; this would present a challenge to the sensitivity of the test kit.

[2] Feldsine et al., AOAC INTERNATIONAL Methods Committee Guidelines for Validation of Qualitative and Quantitative Food Microbiological Official Methods of Analysis. Journal of the AOAC International 85(2): 1187-1200.

[3] AOAC International Presidential Task Force, Best Practices in Microbiological Methodology (August 10, 2006), accessed at: http://www.fda.gov/Food/ScienceResearch/LaboratoryMethods/ucm124900.htm

5. The study should evaluate the test's performance against a cultural method. For most FSIS-regulated products, the current FSIS method, found in the microbiology laboratory guidebook (MLG), is the most appropriate reference cultural method[4].
6. The study should evaluate a sufficient number of samples. The number of samples should be chosen to provide adequate statistical assurance that a false negative conclusion will not be reached (i.e., that **A** and **R** were equivalent when, in reality, the methods were not equivalent).

This guidance document in section II provides a robust validation experimental design that addresses the above mentioned parameters. The document can be used by test kit developers, laboratories, or independent validation organizations to determine whether a new alternative method **A** would be appropriate for testing programs conducted in FSIS-regulated establishments. Two criteria should be considered:

1. Demonstration that recovery rates for **A** and **R** are statistically indistinguishable using an unpaired trial with fractional recovery of positive results[5].
2. Evaluate the sensitivity of **A** using a paired trial and a minimum of 29 positive samples.

FSIS will use these data to evaluate a manufacturer's claim that a new method was equivalent to a reference method, including the FSIS method.

[4] When minimal changes have been introduced, validation against a non-cultural method may be appropriate.

[5] A different study design would be needed to demonstrate that recovery rates for **A** were superior to **R**. This situation is not addressed in this guidance document.

II. General Guidance for Evaluation of Pathogen Test Kit Performance

The following guidance is provided to assist the design of effective validation studies that are likely to meet FSIS' expectations.

1. Purpose, Scope and Audience

This guidance document is intended to assist the design of validation experiments for methods used to detect bacterial pathogens in matrices such as meat, poultry, and egg products, and environmental samples (sponges, swabs, brines). In particular, the document could be used to evaluate the performance of a candidate alternative method for *E. coli* O157:H7 or non-O157 Shiga toxin producing *E. coli* (non-O157 STEC) strains. The document is not applicable to methods for enumerating microorganisms.

This guidance document focuses on procedures to measure sensitivity (and false negative rate), specificity (and false positive rate), and to compare positive recovery rates for an alternative and reference method (abbreviated as **A** and **R**, respectively). These measures of test performance should be evaluated when any modification is introduced to **A**, including, for example, changes to test portion size, enrichment media, enrichment time, enrichment temperature, sample to media ratio, or test matrix. If a major modification is introduced to **A**, then Inclusivity, Exclusivity, Repeatability, Reproducibility and Ruggedness Testing should also be performed, either through the direction of an independent organization, or by following guidance provided by the AOAC International Official Methods of Analysis Program[2]. A major modification to **A** would include significant changes in the design or the component reagents for a screening test, for example, the introduction of a new antibody or oligonucleotide primer.

This guidance document is not intended to conflict with or supplant existing guidance from independent organizations (AOAC and ISO), and is not intended to have any application to the initial validation of HACCP plans described in 9 CFR 417.4(a)(1).

The intended audience for this document includes test kit manufacturers, laboratories as well as independent organizations that evaluate test kit performance.

2. General Considerations

The work should be carried out in a laboratory that is independent of the manufacturer's economic interest. For example, the study may be carried out under contract to an academic laboratory, or a publicly-, or privately-owned laboratory that is not controlled by the test manufacturer. Alternatively, the validation may be performed through an independent organization such as AOAC, AFNOR, ISO, or NordVal. To avoid handling bias, the identity of the samples should be blinded to the analysts. The study design should be reviewed by an outside party before initiating work. FSIS can review and comment on study design[6]. Finally, all study reports as well as the associated raw data should be available for review by FSIS.

[6] Submit proposals for FSIS comment through the sampling queue at askFSIS (http://askfsis.custhelp.com).

3. Inoculum

Number of strains: The number of strains to be used for inclusivity and exclusivity studies is referenced in the AOAC guidelines[2]. Experiments to determine method equivalence should be conducted under conditions that result in fractional recovery of positive samples. Therefore, the use of multiple strain cocktails is not recommended, because individual strains would segregate to different samples.

Strain selection: Strains used to measure test performance should be available from public collections (e.g., ATCC, DSMZ, JCM), academic government reference laboratories, or other collections that are available to the scientific community. For inclusivity and exclusivity studies, the strain set should include strains that do and do not meet the FSIS regulatory definition based on the current FSIS MLG method (Table 1). For validation of test performance characteristics, the target strain should be associated historically with the matrix, or an outbreak. These should be the first strains of choice for conducting the validation study. Strains demonstrating reduced growth potential in particular matrices also should be chosen to challenge the validation. A validation experiment using a challenging target strain would provide additional information on the robustness of the alternative procedure. If FSIS had evidence that certain strains or serotypes consistently were not detected by a commercially available method, it may request additional validation data. Recommendations for typical strains of *E. coli* O157:H7, *Listeria monocytogenes*, *Listeria* species, and *Salmonella* strains are provided in Table 1. FSIS welcomes recommendations for typical and challenging strains to be considered for validation studies in specific meat, poultry, egg product, and environmental matrices.

Inoculum preparation: To insure the purity of the target strain, a single, isolated colony is picked from a non-selective plating medium. For experiments to determine method equivalence, the isolated colony is used to inoculate an appropriate liquid medium, and it is incubated until the culture reaches stationary phase. Following incubation, the stationary phase culture should be cold-stressed (4°C, 18-24 hours)[7]. After 24 hours at 4°C, the culture can be diluted and plated on a non-selective medium to determine colony forming units per milliliter (CFU/mL). Sufficient measurements should be made to determine that the target strain is uniformly distributed in the culture to ensure that the inoculum in the tested samples is distributed as a Poisson distribution. These results should be reported. The 24 hour CFU/mL value can be used to determine the volume of inoculum to be added to the matrix to achieve the desired target strain concentration. Alternatively, the target strain level in the inoculated matrix can be estimated by most probable number (MPN) analysis. FSIS welcomes recommendations for alternative procedures for preparing target strains for validation experiments and for inoculating test matrices.

[7] This is a minimum recommendation for stress conditioning of the inoculum. The study design should consider the typical conditions used to manufacture the matrix of interest at the typical point of sampling. These may include temperature extremes, salt, water activity, pH, or the presence of residual antimicrobial compounds like organic acids. In some cases, the inoculum can be exposed to extreme conditions simply by exposure to the test matrix.

Inoculation of the matrix: The inoculation level should be sufficient to result in fractional recovery of positive samples per test portion, defined as a range of 20-80% confirmed positive results for **R**. The inoculation level merely refers to the average level of target organism delivered to each test portion that would result in fractional recovery of positive results. The level may be different based on the choice of **R** or target organism, and higher inoculation levels may be required if the fractional recovery rate does not meet the recommended range. If possible, the matrix should be well mixed before inoculation to reduce potential variation in composition including intrinsic interfering factors. For example, high fat (e.g., 50% lean) beef trim can be sliced into small pieces or ground to distribute the fat and background microflora before inoculation. The experimental portions should be prepared and inoculated with the necessary volume of inoculum preparation to ensure the inoculum is well distributed. If a multicomponent product is to be evaluated, the non-FSIS and FSIS-regulated components should be likewise well distributed before inoculation. If multiple portion sizes are to be evaluated, a portion of the matrix can be inoculated at X CFU per 25 grams, and then 25 gram portions of inoculated matrix are combined with additional, uninoculated matrix to create alternative portion sizes containing X CFU per portion. For validation of environmental testing methods, the inoculum may be added directly to the collection device (swab or sponge). However, the typical conditions of use should be simulated including the presence of competitive microflora. For example, before adding the inoculum, the device should be used to swab a surface. The device and inoculum should be combined with sample collection media before enrichment. The same concept of low level inoculation and fractional recovery apply to the validation of tests for environmental samples. A number of samples (5-10 per trial) should be uninoculated to serve as negative controls.

4. Matrix

Choice of Matrix: Validation studies should use matrices typical of the samples likely to be tested. Food matrices can be mixed before inoculation and enrichment to reduce potential for experimental variation. The choice of the matrix and the decision to initiate a validation study for a new matrix should be based as much as possible on the intrinsic properties of the matrix that are likely to affect the growth of the target pathogen. These properties include: levels of indigenous microflora, fat content, pH, salt content, water activity, the presence of antimicrobial compounds including additives found in ready-to-eat products, and the presence of residual antimicrobial compounds typically used for treating environmental surfaces or raw products. The intrinsic properties should be evaluated at a location in the process when the sample is likely to be collected. For example, beef trim is typically sampled after fabrication, so a validation study intended for beef trim should use trim collected at that point. Primal cuts purchased at retail would not be a suitable substitute. Similarly, a *Salmonella* test intended for raw egg yolk may not be suitable for testing pasteurized egg yolk product containing additives that may interfere with *Salmonella* growth. A scheme for determining meat, poultry, and egg matrix categories based on water, fat, spice, salt, or sugar content, and cooking was created by the Presidential Task Force on Best Practices in Microbiological Methodology (BPMM)[3]. Additional factors described above should be considered as well. Examples illustrating this concept are found in Box 2. Tables 2a and 2b provide some examples of matrix categories for FSIS-regulated meat, poultry, and egg products.

Box 2. Decision Criteria Illustrating When New Validation Studies Should be Conducted for an Existing Method

- A commercially available test for *E. coli* O157:H7 was AOAC-validated for ground beef products. The test matrix was 80% lean ground beef. Customers would like to use the test for lean beef trim of comparable fat level. There is no need to re-validate the test for this matrix, since there is minimal difference in fat content, an important intrinsic property that may affect test performance. No other intrinsic properties (such as residual antimicrobial compounds) distinguish these matrices.
- Customers want to use the above mentioned test for 50% lean trim, a validation experiment should be designed for this matrix since the fat content is significantly different and may affect test performance.
- A commercially available test for *Listeria monocytogenes* was validated for ready to eat turkey roll with 5% salt added. It was concluded that there is no need to re-validate the test for use with low fat beef hot dogs because the fat and salt content were similar and are not expected to have a differential effect on test performance based on comparative information about growth kinetics within turkey and beef matrices. However, use of the test with dry fermented salami would require additional validation because the reduced water activity and presence of added microbial flora (lactic acid bacteria) in the salami could affect test performance.
- The above mentioned *L. monocytogenes* test should be validated for use with sponge samples collected from environmental surfaces.

Characterization of matrix: Intrinsic properties of concern for the specific matrix (e.g., APC to evaluate microbial flora, water activity, pH, antimicrobial residues or additives) should be measured in the material chosen for the study, and the values should be compared to published or unpublished ranges for the product type. The values (as well as the methods for determining the values) should be presented in the study report or should be otherwise available for review by FSIS.

5. Study Design and Analysis.

Paired and Unpaired studies: Validation studies should measure performance characteristics of an alternative method (**A**) relative to a reference method (**R**). Figure 1 illustrates validation study designs. Microbiological methods typically involve sequential sample preparation and enrichment, screening, and confirmatory procedures. Figure 1 depicts these procedures using the numbers P, S, and C respectively. For example, **AP** refers to the alternative sample preparation and enrichment procedure, while **RP** refers to the sample preparation and enrichment procedures indicated in the reference method. **AC** refers to the reference confirmatory procedure applied to **A**.

Validation studies should rely on two components to determine the equivalency of **A** and **R**: relative recovery and sensitivity (false negative rate). FSIS believes that the recovery **A** and **R** should be statistically indistinguishable, and that a robust estimate of sensitivity should be determined. FSIS has not determined that a specific sensitivity criterion is appropriate for all situations. Manufacturers should evaluate sensitivity estimates on a case by case basis. Test kit manufacturers can use this guidance document to evaluate sensitivity for a test kit, or to demonstrate that a test kit met a specific sensitivity criterion.

Two study designs in Figure 1 illustrate how relative recovery and sensitivity are determined and calculated. An unpaired study design is intended to compare the recoveries of **A** and **R**. It is performed using independent samples that are randomly assigned to either the **A** or **R** procedure[8]. A paired study design is intended to measure the sensitivity of **A**[9]. The paired study is performed by taking two or more measurements from the same sample to which **A** is applied[10].

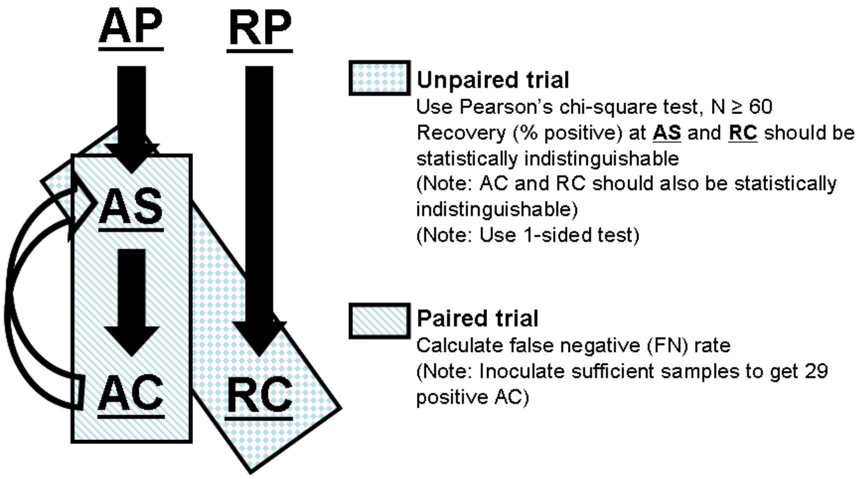
Evidence from both unpaired and paired studies is used to evaluate a manufacturer's claim that a new method was equivalent to a reference cultural method, including the FSIS method.

[8] Care must be taken to avoid biased selection of samples for the A or R protocol.

[9] specificity (false positive rate), can also be estimated using a paired experiment as the ratio of negative tests at **AS** divided by **AC**.

[10] For example, a sample is prepared by the **A** method and is sampled after 15 hours of enrichment with **AS** (the alternative screening test) and **AC** (the reference confirmatory procedure).

Figure 1. Validation Experiment Design. **A** is the alternative method, **R** is the reference method. P is the sample preparation and enrichment procedure, S is the screening procedure, and C is the reference confirmatory procedure. **AC** refers to the reference confirmatory procedure applied to **A**. **RS** refers to a screening procedure applied to **R**. In an unpaired trial, recovery (% positive) at **AS** is compared to recovery at **R2**. Results at **AS** should be confirmed at **AC**, and recovery at **RC** and **AC** should not be statistically distinguishable. In a paired trial, a false negative or false positive rate is calculated as the ratio of positive tests at **AS** divided by **AC**.



▨ **Unpaired trial**
Use Pearson's chi-square test, N ≥ 60
Recovery (% positive) at **AS** and **RC** should be statistically indistinguishable
(Note: AC and RC should also be statistically indistinguishable)
(Note: Use 1-sided test)

▨ **Paired trial**
Calculate false negative (FN) rate
(Note: Inoculate sufficient samples to get 29 positive AC)

Data Analysis. Data analysis from an unpaired study typically involves a comparison of recovery rates (proportion of positive results) at **AS** and **RC**, but only if fractional recovery is achieved. In other words, the proportion of positive samples using **A** is compared to the proportion of confirmed positives using **R**. Because some of the **AS** results may in fact be false, all **AS** results used in this comparison are confirmed (**AC**). A simple Pearson chi-square test without continuity correction should be used to determine if the recovery rates from the two procedures are statistically distinguishable. The associated P-value for significance (alpha) provides the probability that the methods are found not to be equivalent by chance, assuming the true proportions were identical. FSIS' concern is that the performance of **A** should not be inferior to **R**. Therefore, the statistical test should decide between the null hypothesis (that is, the performances of **A** and **R** are identical) and the alternative hypothesis (that the performance of **A** is inferior to **R**). Thus, a 1-sided statistical test should be used. By convention, alpha is set to 0.05 (meaning that there is a 5% probability that the statistical evidence would lead to accepting the alternative hypothesis when in fact the null hypothesis is true). A Pearson chi-square test without correction factors is recommended because this method is commonly used for statistical testing and does not rely on access to, or knowledge of, sophisticated computer programs that may not be available to all[11] (see attachment). The

[11] The Pearson Chi square statistic is made using the familiar $(O-E)^2/E$ formula: where O is the observed result, E is the expected value of the result, assuming the truth of the null hypothesis, summing over the 4 "cells" of a 2 x 2 table that has entries equal to the number of positive and negative results for the two methods. The expected value under the null hypothesis is the average of the two corresponding method specific results. The statistical significance (p- value) of the result of this calculation is equal to the

one-sided chi-square test at an alpha of 0.05 is essentially the same as the two-sided chi-square at an alpha of 0.10. Thus, rejection of the null hypothesis and acceptance of the alternative hypothesis occurs when the chi-square statistic exceeds 2.7055 and the number of positive results for **R** is greater than that for **A**.

Data analysis from a paired study involves a comparison of paired results from the same sample determined by **AS** and **AC**. Sensitivity and specificity are determined from the ratio of **AS** to **AC** for positive results (to determine sensitivity) or for negative results (to determine specificity)[12]. These ratios are point estimates, and unless a large number of samples test positive or negative, the associated confidence intervals would be very wide, and would be used to provide a robust estimate of sensitivity or specificity (see below for discussion of sample size).

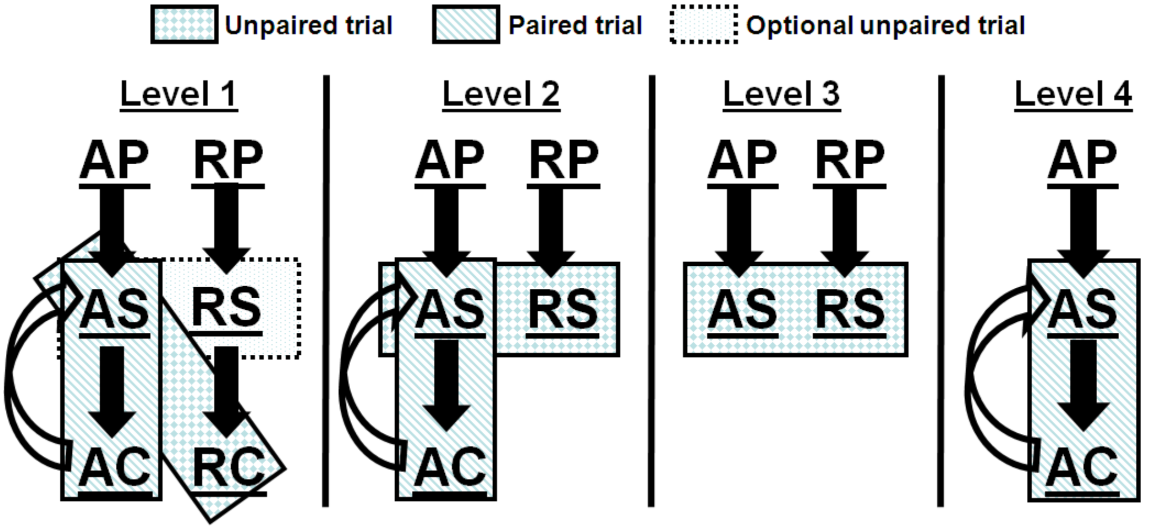
FSIS Levels of validation. FSIS believes that all alternative methods should be validated using robust studies such as those described above. However, the Agency realizes that modifications to methods may not always require the same level of validation. Therefore, FSIS proposes four levels of validation that may be appropriate for some circumstances (Figure 2, Table 3). Note that a minimum of 60 samples per method are recommended for all levels.

- FSIS level 1 validation includes unpaired and paired studies, and all samples are confirmed using the reference confirmatory procedure (**RC** for **R** or **AC** for **A**). Level 1 is recommended for any novel **A**, or when a major modification, or two or more non-major modifications are introduced to **A** (e.g., new matrix, new screening test, or new enrichment broth). Comparison of recovery of the screening device with the **A** and **R** methods is optional.
- FSIS level 2 validation could be used when a single non-major modification is made to **A**. Like level 1, level 2 also includes an unpaired and paired study, but would allow a screening test to substitute for the full reference method. This screening test (referred to as **RS**) would be recommended only if the performance is determined using a level 1 validation. In this situation, the "apparent" equivalency of **A** is determined by comparing recovery rates for **AS** with **RS**
- FSIS level 3 validation could also be used when a single non-major modification is made to **A**. Level 3 only includes an unpaired study, and would allow substitution of a screening test (**AS** and **RS**) for confirmed results provided that the same screening test was used for **A** and **R,** and the performance of **AS** and **RS** were determined using level 1 validation. Level 2 is preferable to level 3 validation.
- FSIS level 4 validation would be appropriate only when a full reference method does not exist. Level 4 only includes a paired study in which recovery at **AC** and **AR** is compared.

probability that a chi-square-distributed random variable (with 1 degree of freedom) exceeds the result. The attachment describes how the Pearson chi-square test statistic is calculated.

[12] Unlike the paired experiment, **AS** results are not corrected by **AC** results.

Figure 2. FSIS levels of Validation. **A** is the alternative method, **R** is the reference method. P is the sample preparation and enrichment procedure, S is the screening procedure, and C is the reference confirmatory procedure. **AC** refers to the reference confirmatory procedure applied to **A**. **RS** refers to an alternative screening procedure applied to **R**. In an *unpaired trial*, recovery (% positive) at **AS** is compared to recovery at **R2**. Results at **AS** should be confirmed at **AC**, and recovery at **RC** and **AC** should not be statistically distinguishable. In a *paired trial*, a false negative or false positive rate is calculated by comparing results at **AS** and **AC**. Comparison of recovery by **RS** and **AS** is optional for FSIS level 1 validation.



Reference method: For FSIS regulated products, the current FSIS method, which is found in the Microbiology Laboratory Guidebook (MLG), is the most appropriate reference cultural method for validating methods used by FSIS-regulated establishments. Other cultural methods also may be appropriate, including methods described in FDA's Bacterial Analytical Manual (BAM), or reference methods defined by the International Standards Organization (ISO), or the *Codex Alimentarius*. In certain circumstances, as indicated in Table 3 and figure 2, a non-cultural method may be appropriate for validation studies when minimal changes have been introduced to **A**, and the non-cultural method is well-defined (see Levels of Validation). Alternative methods should be re-validated when significant changes affecting performance are introduced to the reference method. Re-validation should be performed within one year of the introduced changes.

6. Sample size

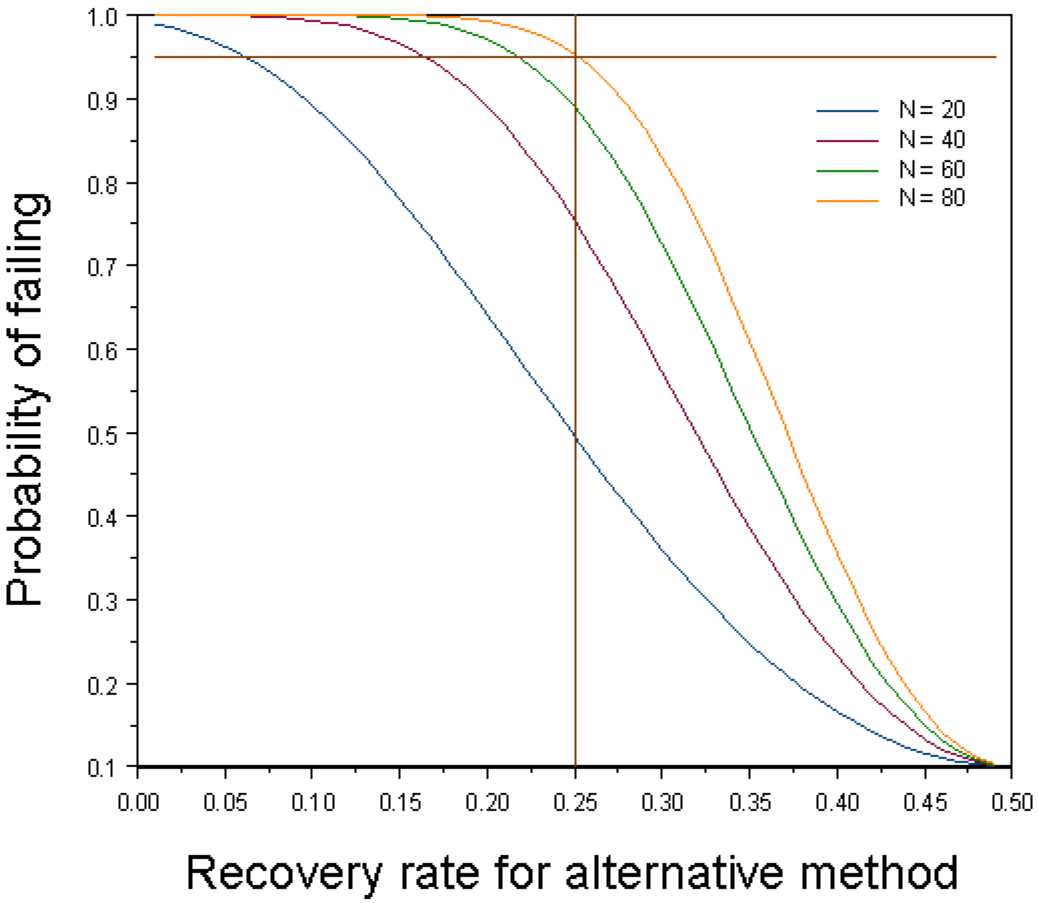
To provide robust estimates of method equivalency, as well as estimates of sensitivity and specificity, sample size needs to be addressed.

Unpaired studies: The number of samples tested per method, the anticipated recovery rates, and the confidence level (alpha), are used to calculate statistical power, defined as the probability of detecting a true difference between **A** and **R**. That is, the number of samples that would provide high probability of detecting a difference in the recovery rates of **A** and **R**. FSIS calculated statistical power for experiments with different sample sizes and underlying differences in recovery rate[13]**.** In these calculations, FSIS assumed alpha to be 0.05, and a 50% recovery rate for **R**. Figure 3 presents calculated statistical power as a function of **A** and **R** recovery rates and sample size. FSIS would not accept as equivalent a candidate **A** performing at 50% relative to R. In other words, the recovery of **R** is 50% and that of **A** is 25%. Thus, in this scenario, there should be a high probability of rejecting the null hypothesis and accepting the alternative hypothesis. For an experiment with N = 20 (i.e., 20 samples tested per method), the power is 49.5%. That is, in over one-half of the experiments, the null hypothesis would not be rejected and the recovery rate for **A** and **R** would be judged as indistinguishable. If N = 40, the power is 75.4%, an almost 25% chance of not detecting what FSIS would consider to be a large difference between the recovery rates of **A** and **R**. When N = 60, the power is 89%, almost 90%; that is, an almost 9:1 odds of detecting such a true difference; for N = 80, there is a 95% power, that is, about a 19:1 chance of detecting the difference.

For these reasons, FSIS recommends a minimal sample size of 60 per method (and preferably 80 samples per method) to determine robustly if the recovery rates of **A** and **R** are indistinguishable.

[13] Using PROC POWER procedure of SAS® version 9.1

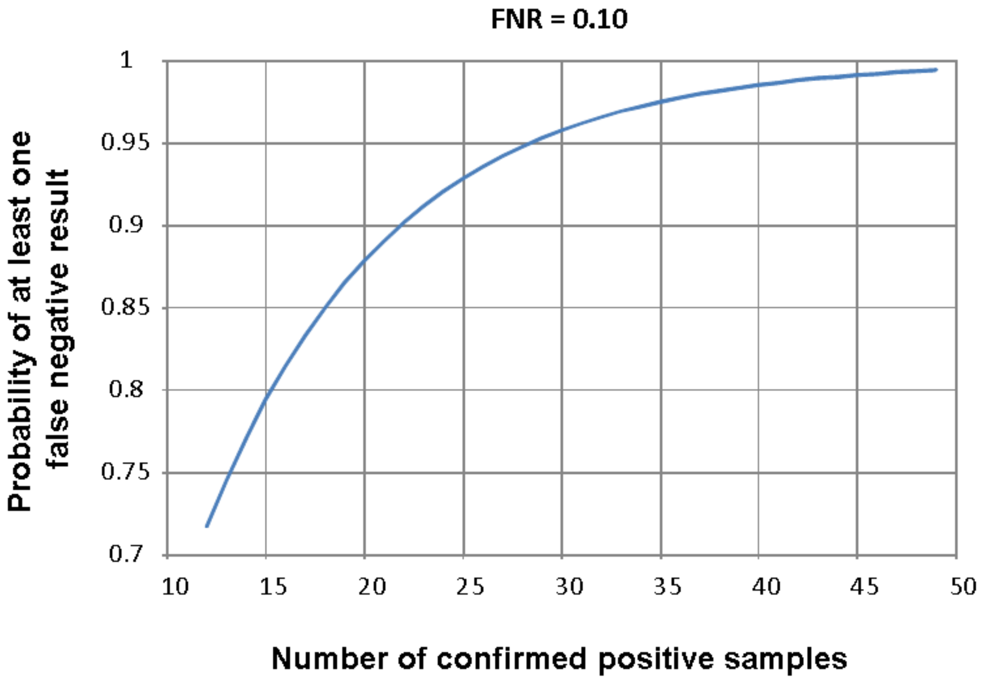
Figure 3. Statistical Power as Determined by Sample Size, Recovery Rate, and Minimal Difference between the Alternative and Reference Methods. Statistical power was calculated using PROC POWER program of PC SAS® version 9.13, assuming 50% recovery rate for the reference method, 10% significance (i.e., a 1-sided test with 5% significance), and 5-45% reduced recovery for the alternative method compared to the reference method. The probability of determining that the methods are statistically distinguishable is shown.



Paired studies: As indicated above, a large number of samples would be needed to provide a robust estimate of sensitivity. FSIS recommends that validation studies include a large number of paired trials of **A** that confirm positive by **AC**. Figure 4 shows the probability of finding at least one false negative result among 12 to 50 confirmed positive samples assuming a hypothetical alternative test with 90% sensitivity (i.e., a 10% false negative rate). There should be a high level of assurance that the validation study would detect at least one false negative result from this hypothetical test. The curve indicates that 29 or greater positive samples would provide high (95% or higher) assurance that at least one false negative result would be detected from the hypothetical test. Zero false negative results from 29 confirmed positives would be consistent with a test having a sensitivity that met or exceeded 90% and zero negative results from 50 confirmed positives would be consistent with a test with a sensitivity that met or exceeded 94%.

These calculations are provided for illustrative purposes; FSIS has not determined that a specific sensitivity criterion is appropriate for all situations. The key point is that a larger number of confirmed positive samples provide greater assurance of detecting an unacceptable false negative rate.

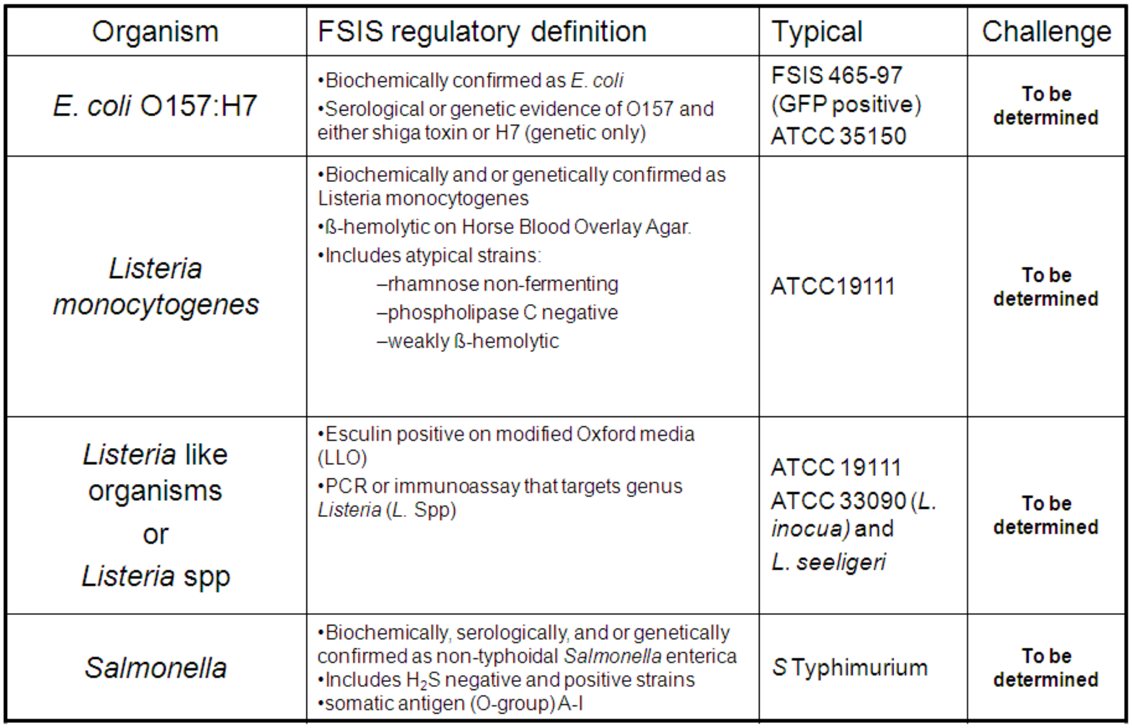
Figure 4. Probability of Finding at Least One False Negative Result for a Hypothetical Test with a 10% False Negative Rate Versus Number of Samples. Twenty-nine samples provide 95% probability (confidence) of detecting at least 1 false negative with this test.



7. Study Report

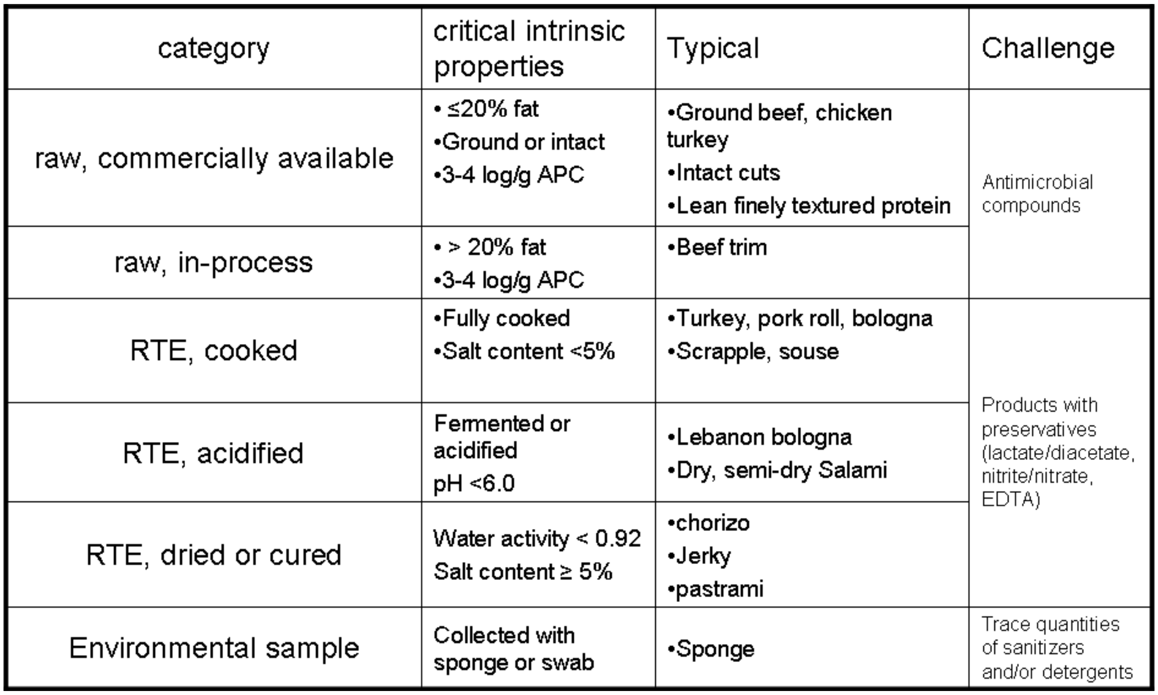
Preferably, validation study reports should be published in an appropriate peer-reviewed journal, such as the *Journal of the AOAC International*. In any case, a study report containing experimental details and format similar to a scientific article format should be provided to FSIS for review. This would include abstract, introduction, materials and methods, results, discussion, and references sections. For new methods or modifications to existing methods that have not been validated by a recognized independent body, a study report and all associated raw data should be available to FSIS for review. Any recommended changes to the validated test protocol should be communicated as soon as possible to new and existing end users as part of a package insert, on the manufacturer’s web site, and in the manufacturer’s technical literature.

Table 1. Typical and Challenging Strains for use with Validation Studies.



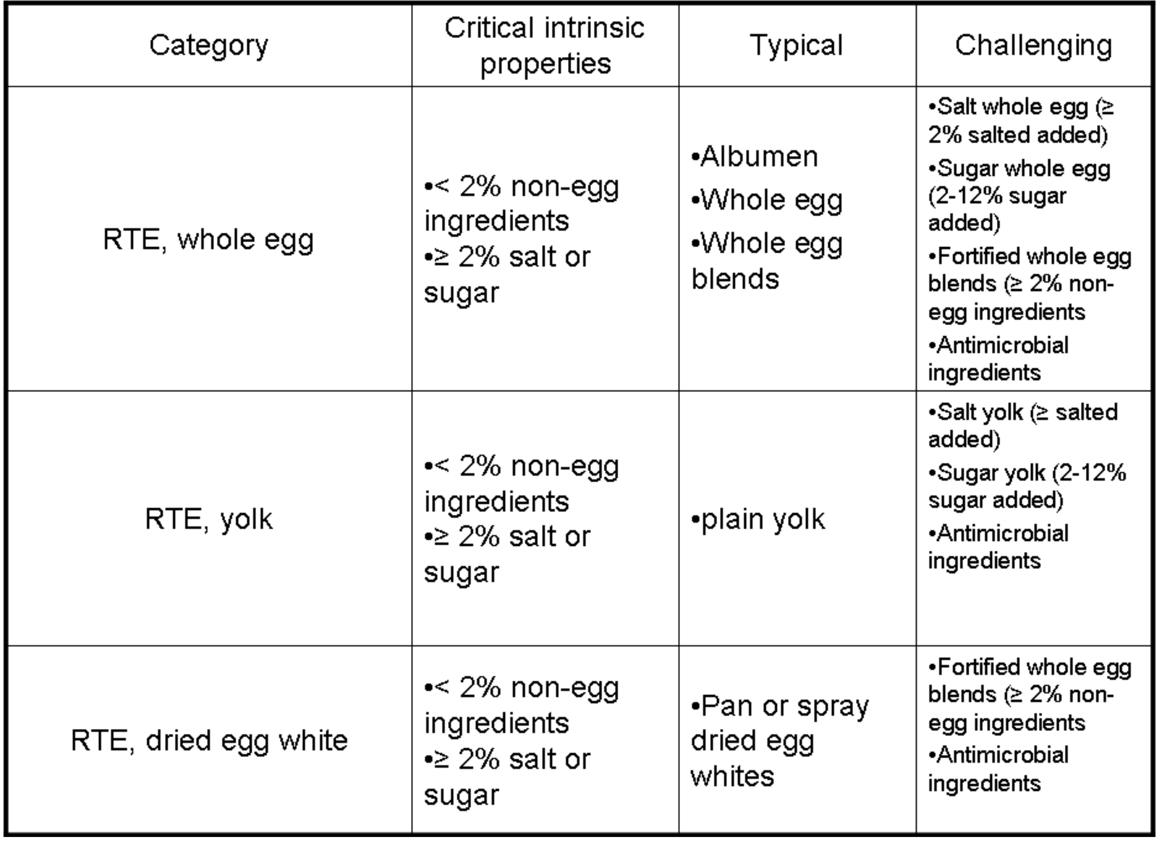
Organism	FSIS regulatory definition	Typical	Challenge
E. coli O157:H7	•Biochemically confirmed as *E. coli* •Serological or genetic evidence of O157 and either shiga toxin or H7 (genetic only)	FSIS 465-97 (GFP positive) ATCC 35150	**To be determined**
Listeria monocytogenes	•Biochemically and or genetically confirmed as Listeria monocytogenes •ß-hemolytic on Horse Blood Overlay Agar. •Includes atypical strains: –rhamnose non-fermenting –phospholipase C negative –weakly ß-hemolytic	ATCC19111	**To be determined**
Listeria like organisms or *Listeria* spp	•Esculin positive on modified Oxford media (LLO) •PCR or immunoassay that targets genus *Listeria* (*L.* Spp)	ATCC 19111 ATCC 33090 (*L. inocua)* and *L. seeligeri*	**To be determined**
Salmonella	•Biochemically, serologically, and or genetically confirmed as non-typhoidal *Salmonella* enterica •Includes H_2S negative and positive strains •somatic antigen (O-group) A-I	*S* Typhimurium	**To be determined**

Table 2a. Typical and Challenging Meat and Poultry Matrices for use with Validation Studies.



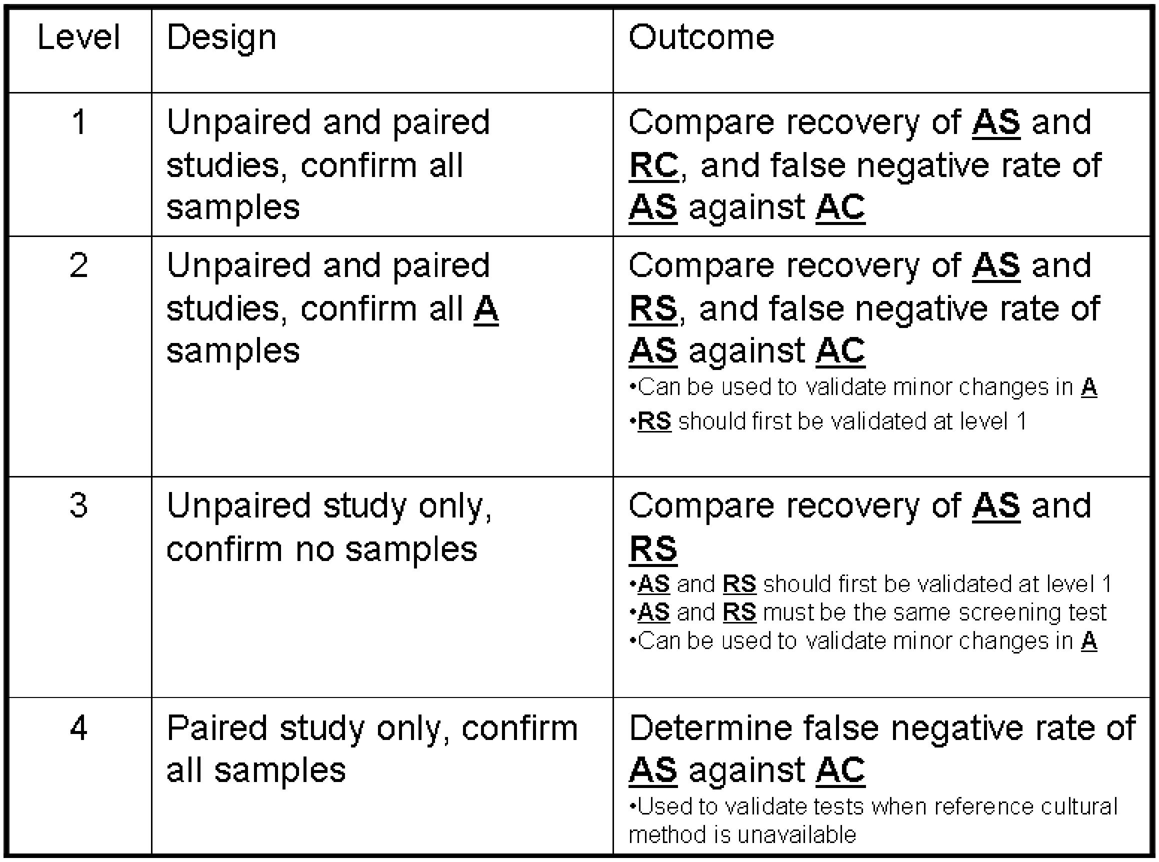
category	critical intrinsic properties	Typical	Challenge
raw, commercially available	• ≤20% fat •Ground or intact •3-4 log/g APC	•Ground beef, chicken turkey •Intact cuts •Lean finely textured protein	Antimicrobial compounds
raw, in-process	• > 20% fat •3-4 log/g APC	•Beef trim	
RTE, cooked	•Fully cooked •Salt content <5%	•Turkey, pork roll, bologna •Scrapple, souse	Products with preservatives (lactate/diacetate, nitrite/nitrate, EDTA)
RTE, acidified	Fermented or acidified pH <6.0	•Lebanon bologna •Dry, semi-dry Salami	
RTE, dried or cured	Water activity < 0.92 Salt content ≥ 5%	•chorizo •Jerky •pastrami	
Environmental sample	Collected with sponge or swab	•Sponge	Trace quantities of sanitizers and/or detergents

Table 2b. Typical and Challenging Egg Product Matrices for use with Validation Studies



Category	Critical intrinsic properties	Typical	Challenging
RTE, whole egg	•< 2% non-egg ingredients •≥ 2% salt or sugar	•Albumen •Whole egg •Whole egg blends	•Salt whole egg (≥ 2% salted added) •Sugar whole egg (2-12% sugar added) •Fortified whole egg blends (≥ 2% non-egg ingredients •Antimicrobial ingredients
RTE, yolk	•< 2% non-egg ingredients •≥ 2% salt or sugar	•plain yolk	•Salt yolk (≥ salted added) •Sugar yolk (2-12% sugar added) •Antimicrobial ingredients
RTE, dried egg white	•< 2% non-egg ingredients •≥ 2% salt or sugar	•Pan or spray dried egg whites	•Fortified whole egg blends (≥ 2% non-egg ingredients •Antimicrobial ingredients

Table 3. FSIS Levels of Validation. **A** is the alternative method, **R** is the reference method. S is the screening procedure, and C is the reference confirmatory procedure. **AC** refers to the reference confirmatory procedure applied to **A**. **RS** refers to an alternative screening procedure applied to **R**. In an *unpaired trial*, recovery (% positive) at **AS** is compared to recovery at **R2**. Results at **AS** should be confirmed at **AC**, and recovery at **RC** and **AC** should not be statistically distinguishable. In a *paired trial*, a false negative or false positive rate is calculated by comparing results at **AS** and **AC**.



Level	Design	Outcome
1	Unpaired and paired studies, confirm all samples	Compare recovery of **AS** and **RC**, and false negative rate of **AS** against **AC**
2	Unpaired and paired studies, confirm all **A** samples	Compare recovery of **AS** and **RS**, and false negative rate of **AS** against **AC** •Can be used to validate minor changes in **A** •**RS** should first be validated at level 1
3	Unpaired study only, confirm no samples	Compare recovery of **AS** and **RS** •**AS** and **RS** should first be validated at level 1 •**AS** and **RS** must be the same screening test •Can be used to validate minor changes in **A**
4	Paired study only, confirm all samples	Determine false negative rate of **AS** against **AC** •Used to validate tests when reference cultural method is unavailable

Attachment: Example of Pearson Chi-square Statistic Calculation for Unpaired Samples

The results of a fractional recovery experiment can be given in a table, represented in Table A-1.

Table A-1: Representation of Results A, B, C, and D from a Fractional Recovery experiment.

	Alternative	Reference
Positive	A	C
Negative	B	D
Percent Positive	100%A/(A+B)	100%C/(C+D)

For example, suppose 60 samples are inoculated at fractional recovery and tested using the reference method (**R**) and 46 are determined to be positive using the reference confirmatory procedure (**RC**). Another 60 samples are inoculated and tested with the alternative method (**A**), and 37 are determined to be positive using the alternative screening test (**AS**) and are confirmed subsequently using the reference confirmatory procedure applied to the alternative method (**AC**). Table A-2 depicts the hypothetical results.

Table A-2: Hypothetical Results of a Fractional Recovery Experiment.

	Alternative	Reference
Positive	37	46
Negative	23	14
Percent Positive	61.7%	76.7%

Calculation: The Pearson chi-square test statistic formula is popularly known as the sum of terms $(O-E)^2/E$, where O is the observed number of results in the cell; E is the expected number of results in the cell when the null hypothesis is true. The sum of these terms over all cells of the table that contain the results (Table A2) gives the value of the chi-square statistic.

The expected numbers of results in the cells, E, are calculated under the assumption that the null hypothesis is true. The null hypothesis states that the recovery of positive results by the two methods is the same. The “expected” numbers of positive results in the four cells are shown in the following table as expected cell values E(A), E(B), E(C), and E(D), if two procedures have the same fractional recoveries.

Table A-3: Expected Number of Cell-Specific Results, E, used in the Calculation of the Chi-Square test, based on Table A1-1.

	Alternative	Reference
Positive	(A+C)/2	(A+C)/2
Negative	(B + D)/2	(B +D)/2
Percent Positive	100%(A+C)/(A+B+C+D)	100%(A+C)/(A+B+C+D)

In the example given above the table of expected values are given in Table A-4.

Table A-4: Expected Number of Cell-Specific Results, E, based on Actual Results in the Example, given in Table A1-2.

	Alternative	Reference
Positive	41.5	41.5
Negative	18.5	18.5
Percent Positive	69.2%	69.2

The chi-square statistic is computed as the sum of the terms $(O\text{-}E)^2/E$ over the four cells of the numbers of results, as shown in formula (1).

Formula (1): $\text{Chi-sq} = (A\text{-}E(A))^2/E(A) + B\text{-}E(B))^2/E(B) + C\text{-}E(C))^2/E(C) + D\text{-}E(D))^2/E(D)$

For the example:

$$\text{Chi-sq} = 0.48795 + 1.09459 + 0.48795 + 1.09459 = 3.16509$$

Note: Because of the symmetry of the calculations for **A** and **R**, it is only necessary to compute the first two terms. A simpler equation can be used, as shown in formula (2).

Formula (2): $\text{Chi-sq} = (A\text{-}C)^2/(A+C) + (B\text{-}D)^2/(B+D)$ (2)

For the example:

$$\text{Chi-sq} = (9)^2/83 + (9)^2/37 = 0.9759 + 2.1892 = 3.1651$$

The Pearson chi-square result of 3.1651 is greater than the cut-off value and the number of positive results for **R** is greater than that for **A**. Thus, the null hypothesis is rejected, and thus, based on these results the recovery of **A** and **R** cannot be considered the same.