

Assay Validation Methods

This document contains educational materials developed by Dr. Larry Hammell, and is used with permission. This information was adapted from sections of the 'Applied Aquaculture Epidemiology' course, developed by the Atlantic Veterinary College and Canadian Aquaculture Institute. The USFWS acknowledges and thanks Dr. Hammell for the use of this material and his contributions to quality assurance in applied aquaculture.

Comparable Testing Methods – Measuring Agreement

The true disease status is frequently unknown, or impossible to obtain with reasonable effort and costs. In many cases, we use imperfect tests for which there is no quantitative measurement of sensitivity or specificity. Even when we spike sample sets with known pathogens, we cannot be sure these sample sets mimic what occurs in a natural infection.

When new technology provides new methodologies, the new test is often compared to the standard testing methods already in practice. Most frequently, the test producing the greatest number of positives is chosen, and assumed to be the best representative of the actual number of positive individuals in the population. This seems to make sense, but from our previous discussion, we know that if tests produce a disproportionate number of false positives, or false negatives.

Often, when two tests are compared, and the total number positive is similar, say for TEST A and TEST B, it is assumed these are the same individuals testing positive in each test. Often, the positive tests on TEST A may be different individuals than the positive tests on TEST B. When this is the case, it may be difficult determining which disease-positive individuals are testing positive. Another assessment that can be done when comparing two tests is to examine the extent of agreement between the tests, taking into consideration the fact that some individuals will test positive on both tests due to chance alone. Let's compare a bacterial culture test (CULTURE) to an immunological assay (ELISA) in a hypothetical comparison of two tests. Our population of 1000 is tested and the two tests produce these results:

Table 1. Comparison of two tests and measure of agreement

	Standard Test (ST) →			
New Test (NT) ↓	ST +	ST -	Total ST	ST Apparent Prevalence
NT +	99	501	600	60% (.6)
NT -	1	399	400	
Total NT	100	900	1000 (n)	
NT Apparent Prevalence	10% (.1)			

In this example, the observed agreement is 99 (positives) and 399 (negatives) = 498/1000, or 49.8%. This seems reasonable; however we should take into account the agreement that would occur by chance alone.

The **probability of both tests being positive** is the product of the two apparent prevalences:

$$0.10 \times 0.60 = .06$$

The **probability of both tests being negative** is the product of 1 minus the two apparent prevalences:

$$0.4 \times 0.9 = .36$$

The sum of these probabilities is the level of **agreement by chance alone**:

$$.06 + .36 = 0.42, \text{ or } 42\%$$

The **agreement beyond chance** is the observed agreement minus the chance agreement:

$$.498 - .420 = .078, \text{ or } 7.8\%$$

The **maximum level of agreement beyond chance** is 1 minus the chance agreement:

$$1 - .42 = 0.58, \text{ or } 58\%$$

The quotient is called *kappa* – the agreement beyond chance divided by the maximum chance agreement:

$$.078 / 0.58 = .13$$

No agreement beyond chance gives a kappa of zero, and perfect agreement if 1.
A moderate level of agreement is considered when kappa is greater than 0.4 – 0.5