

**Temperament Testing --
Deciding which animals live and which animals die. Science? Gut? Or both?
November 2008**

Making life and death decisions for animals on the doorstep of the animal welfare industry is little different than holding a mirror to our humanity.

In an article on this subject, *Temperament Testing in the Age of No Kill*

<http://www.nokilladvocacycenter.org/pdf/Temperament%20Testing.pdf> Nathan Winograd states, "In order to be fair, a temperament test must do two things: (1) screen out aggression and (2) ensure that friendly, scared, shy, sick or injured dogs do not get wrongly executed. By focusing on the first prong, traditional shelters have ignored the second, a violation that goes to the core of the no-kill ideal: Animals are to be judged and treated as individuals."

Doggone Connection www.doggoneconnection.com is a magnet behavior training program for people having dogs with problematic behaviors who have been shuttled out the back doors of shelters because they either failed, or predictably would have failed the "temperament test." Our trainers are frequently revisiting the temperament testing discussion in relation to the reasons for many of the "failed" dogs who attend our program. Recently I came upon an article in "Bark" magazine, Dog is in the Details by Barbara Robertson http://www.thebark.com/ezine/features_specialFeatures/specialFeatures_04.html. I shared this article with the Doggone Connection trainers and Eileen, one of our top notch trainers, was generous to summarize the collective feedback:

This is one of the better articles that I have seen on "temperament testing" or behavior evaluations in dogs. After reading the article, I several thoughts came to mind.

One of the things that I think gets left out of the discussions of "temperament testing" or "behavior evaluations" for dogs is an understanding of the science of testing in general. That is, testing is used in many contexts such as: personality tests are used to predict job success, running speed to predict base running success, mammograms to predict the development of breast cancer. Just because a test is widely used doesn't necessarily make it appropriate; but, it does seem to be worth noting that the issues raised by people regarding behavior evaluations/temperament tests in dogs are similar to those being wrestled with by researchers and practitioners in other realms as well. So what are the common elements across these circumstances?

First, there are flaws with each and every test so you will always end up with the situation whereby the test "fails" in predicting for the outcome for a given individual. That means you will have a dog that, based on how he tested at the shelter, seems problematic and then goes on to find the right home and is a complete peach. College admissions officers experience a parallel situation when they fail to admit a student who didn't do well on the SAT and the student goes onto to be a great scientist. Or consider the medical case of a woman whose mammogram indicates that she is at low risk for developing breast cancer but then develops a rapidly developing form of this very cancer.

A given test can yield results in any of four categories (for more details see: http://en.wikipedia.org/wiki/Sensitivity_and_specificity). For the sake of an example, I'll use the example of dogs being evaluated at a shelter and will use the admittedly crude categories of "healthy" and "unhealthy" to indicate medical and/or behavioral conditions. An evaluation could produce the following results:

- A. True positive: Unhealthy dogs correctly diagnosed as unhealthy
- B. True negative: Healthy dogs correctly identified as healthy
- C. False positive: Healthy dogs wrongly identified as unhealthy
- D. False negative: Unhealthy dogs wrongly identified as healthy

The “perfect” test would put correctly identified dogs into categories A or B (although none of us is happy when a dog is unhealthy as indicated in category A). But, all tests will be wrong some of the time (either through false positive or false negative). A test that was scientifically designed to be very sensitive will occasionally error and produce results that fall into category C. In this case, a dog may be euthanized because he was considered unadoptable. Alternatively, a test that is very crude will occasionally produce results associated with category D above. In this instance, the result may be that a family leaves a shelter with a dog mistakenly thinking the dog is healthy. Serious harm can occur to the family and/or the dog in this instance as well or the dog may be returned to the shelter once the condition becomes apparent.

Does the fact that all tests will periodically make inaccurate predictions mean that the test completely lacks any value? In my opinion, it doesn't. I'm always a bit surprised when someone makes a blanket statement along the lines of as "I don't believe in temperament tests because my dog would (or did) fail one but he has turned out to be an awesome dog" (category C). Some tests do a better job than others in predicting future outcomes. So it seems to me that the questions should be: Is there a good test for my situation? If there are multiple tests available, is one stronger than another in helping me understand what I'm trying to figure out? What are the strengths and limitations of the test(s)? Could the test help me make decisions about what circumstances this dog be suited or not well suited to handle? Is there information that the test doesn't provide such that I should consider from other sources?

The development of reliable and valid testing tools in any realm is one that takes an enormous amount of patience and discipline (no, I never pursued this line of work for the above stated reason). Great patience is needed because the bar is set extremely high for a test to be considered of any value. That is, as a minimum, tests should provide roughly the same results regardless of who administers it (referred to as inter-tester reliability), it should get roughly the same results from Time 1 to Time 2 (referred to test-retest reliability) and it should have predictive power (referred to as predictive validity). So, a process that looks like it should take an afternoon or at most a week or two is actually quite a bit more complex. Thus, it is not surprising that the tests that have been developed have been long in the making and yet still have significant improvements to be made.

It seems worth mentioning that many people eschew any type of formalized testing of dogs in favor of what their “gut” instinct is telling them. This is a compelling argument in part because it speaks to the idea that the “whole is greater than the sum of the parts” or that we cannot fully know another being simply by totaling up the numbers on a score pad. It seems that many important decisions in life are informed by what our gut instincts tell us. Scouts for professional sports, police officers and from varied walks of life about the “X factor” that makes them choose one person over another as the most likely to be successful or in the case of police officers, which person they think is worthy of investigating further (note: see the work of Malcolm Gladwell in his book *Blink* for a discussion of intuition and decision making). Nonetheless, using gut instincts alone is not without drawbacks as some people have more reliable gut instincts than others and our own gut instincts are subject to enormous variations from one day to the next. In recognition of the role of our gut instinct in its role in decision-making, many of the better formal assessments provide room for this type of information to be taken into account.

So, guess what? Any evaluations (formal or those done by gut instinct) that are done by humans will be imperfect. We are, after all, only human. In the end, I think that our collective efforts toward making successful adoptions and moving to no kill are going to be facilitated when we more fully understand what types of information is available to us and are able to synthesize information from a variety of sources.

Eileen