Research Study:
Comparison of SAFER behavior assessment results in shelter dogs at intake and after a 3-day acclimation period

Study results compiled by National Canine Research Council

Article Citation:
Bennett, S. L., Weng, H., Walker, S. L., Placer, M. & Litster, A. (2015). Comparison of SAFER behavior assessment results in shelter dogs at intake and after a 3-day acclimation period. Journal of Applied Animal Welfare Science, 18(2), 153-168. doi: 10.1080/10888705.2014.999916

Summary and Analysis:
This article is included as one of the few studies that examined test-retest reliability. Based on previous studies that showed changes in stress over time in shelters, the researchers hypothesized that behavior evaluation results would reflect parallel changes. If so, this would beg the question of which results are more representative of the animal's typical and likely behavior in a home and would have overall negative implications for the validity of behavior testing. Bennett, Weng, Placer, and Litster (2015) investigated this possibility by administering the Meet Your Match Safety Assessment for Evaluating Rehoming (SAFER) behavior evaluation twice to 33 shelter dogs.

The authors chose to use SAFER because it is commonly used in shelters throughout the United States. Interestingly, it has not been validated in the peer-reviewed literature, but because of its association with the ASPCA it is well regarded and widely implemented. This choice bolsters the study's external validity.

Seven subtests were administered. First, the experimenter held the dog's head and gazed into its eyes ("Look). The second test involved gently grasping fur and skin along the dog's body ("Sensitivity"). The third subtest was an attempt to initiate play by speaking excitedly and lightly poking the dog ("Tag"). During the fourth subtest the evaluator said, "squeeze" and then gently squeezed the dog's leg and paw ("Squeeze"). This was repeated to see if the dog would respond to the vocal cue. The fifth and sixth subtests used a plastic hand to take away food and toy items, respectively ("Food Behavior" and "Toy Behavior"). Finally, in the seventh subtest the subject was lead into a room occupied by a second, passive dog. Initial approach behavior was recorded, but the dogs were not allowed to touch or interact further ("Dog to Dog Behavior").

The same certified evaluator was used on both days, as was the same assessor. The same helper dog was used on day 0 and day 3 for half (17) of the dogs tested. The remaining 16 dogs experienced different dogs on days 0 and 3.

For each subtest, dogs received a score from 1-5, with higher numbers indicating escalating aggressive behaviors. Specific behaviors were not listed for every score, but a 3 might indicate "signs of fear, high arousal, or inhibited aggression," and a 5 includes growling, lunging, or attempting to bite. According to the assessment's creator, a score of 3 should be interpreted as a recommendation that the dog might benefit from behavior management or modification. For a

more detailed description of the scoring procedure, see the summary and analysis of Bennett et al., 2012).

When analyzing the data, the researchers were particularly interested in cases where scores changed by at least 2 points from day 0 to day 3; these differences were used to calculate percent discordance, or the percent of the sample for each subtest in which scores changed at least 2 points between the two tests. They felt this was of practical importance because differences of this magnitude could conceivably result in different recommendations and vastly different outcomes for dogs (i.e., life or death).

There was at best moderate agreement for 3 out of 4 tests' results studied between days 0 and 3 for this sample. There was little agreement between days 0 and 3 for the first subtest ("Look"); discordance was 15% and weighted kappa was 0.28. Moderate agreement was found for the "Sensitivity" and "Tag" subtests (4% discordance for both, and kappa equal to 0.59 and 0.41, respectively). The first "Squeeze" test showed poor to moderate agreement (8% discordance, kappa of 0.22) while the second squeeze test showed no discordance (kappa = 0.78). Ninety-two percent of the tested dogs scored a 1 or 2 for both "Squeeze" assessments. More than half of the subjects did not have data on both days for the "Food Behavior" subtest due to lack of interest in the food. For the dogs that did have both data points, agreement was poor; discordance was 18% and kappa was 0.50. There was excellent agreement (no discordance) for the "Toy Behavior" subtest with all dogs (except one) scoring a 1 for both assessments. The remaining dog scored a 1 and a 2 on day 0 and day 3, respectively. Finally, moderate agreement across assessments was reported for the "Dog to dog Behavior" subtest; discordance was 3% and kappa = 0.33.

It is interesting to note that for the subtest with the highest agreement ("Toy Behavior"), the topography of behavior was on the lower end of the scale. On higher behavior scores, discordance increased and kappa decreased. Moreover, when behavior did change between assessments, it did not change in a consistent direction. For example, for the "Look" subtest, there were two dogs who scored 5 on day 0 but a score of 1 and 2 on day 3, and there were two dogs who scored 2 on day 0 but had a score of 5 on day 3. There is no clear directional change in behavior.

The most important finding from this study is that even over as short a time period as 3 days, dogs' behaviors can change drastically which may result in vastly different recommendations when these are based on a behavior assessment. This points to a fundamental lack of reliability and external validity for this behavior assessment, which again raises the question of whether these types of evaluations should be used to determine a dog's fate. The authors recommend avoiding testing when dogs are particularly stressed as well as seriously considering the dog's welfare when determining the time of test, and they attributed the differences in results to changes in stress level due to acclimatization to the shelter environment. However nothing in the data demonstrates that the change was anything other than simple unreliability of the test.

Abstract and Link to Purchase Full Text of the Original Article:
https://www.ncbi.nlm.nih.gov/pubmed/25603466