Intra-rater and Inter-rater Reliability of the ASPCA’s Behavior Evaluation of Fearful Dogs

Sara Matters
CUNY Hunter College

How does access to this work benefit you? Let us know!

Follow this and additional works at: http://academicworks.cuny.edu/hc_sas_etds

Part of the Other Psychology Commons

Recommended Citation
Matters, Sara, "Intra-rater and Inter-rater Reliability of the ASPCA’s Behavior Evaluation of Fearful Dogs" (2016). CUNY Academic Works.
http://academicworks.cuny.edu/hc_sas_etds/90

This Thesis is brought to you for free and open access by the Hunter College at CUNY Academic Works. It has been accepted for inclusion in School of Arts & Sciences Theses by an authorized administrator of CUNY Academic Works. For more information, please contact AcademicWorks@cuny.edu.
Intra-rater and Inter-rater Reliability of the ASPCA’s Behavior Evaluation of Fearful Dogs

By

Sara Matters

Submitted in partial fulfillment of the requirements for the degree of Master of Arts, Animal Behavior and Conservation, Hunter College of the City of New York

2016

Thesis Sponsor:

5/16/2016
Dr. James Gordon
Date
First Reader

5/16/2016
Dr. Katherine Miller
Date
Second Reader

5/16/2016
Dr. Martin Chodorow
Date
Third Reader
Abstract

Animal shelter staff commonly assess shelter dogs’ behavior using battery assessments that test the dogs in relation to stimuli they may encounter after adoption. Researchers examining the efficacy of such evaluations typically have not reported on the reliability of the evaluators implementing them. This study examined the intra-rater and inter-rater reliability of four experienced raters using the ASPCA Behavioral Rehabilitation Center Behavior Evaluation of fearful dogs. The raters each watched video recordings of the evaluations of 12 dogs on two separate occasions. They rated the dogs’ behaviors in 21 contexts, then used that information to give the dogs a letter grade (A through D) in seven behavioral areas and one overall letter grade. The ratings were analyzed for intra-rater reliability for each rater and inter-rater reliability among all four raters using measures of agreement, unweighted Kappa, quadratic weighted Kappa and Pearson $r$ statistics. It was predicted that the two raters who used the behavior evaluation regularly would have higher intra- and inter-rater reliability than the two raters who did not. It was also predicted that some subtests would have unacceptably low inter-rater reliability. Both intra-rater and inter-rater agreement and reliability statistics were generally good to excellent across raters and rater pairs. The results of Friedman and Wilcoxon signed ranks test comparisons suggested significant differences among raters and rater pairs. Study hypotheses were generally supported. Results indicated that frequency of use of the behavior evaluation may be more important than years of experience as behavior evaluators in achieving high intra- and inter-rater reliability. Other possible factors affecting rater reliability—watching video recordings of evaluations, rater familiarity with the dogs, and precision of the evaluation in prescribing how some subtests should be rated—are also discussed.

Keywords: Intra-rater reliability, inter-rater reliability, canine behavior evaluation, ASPCA, fear
Intra-rater and Inter-rater Reliability of the ASPCA’s Behavior Evaluation of Fearful Dogs

Animal shelter staff commonly assess shelter dogs’ behavior to screen for behavior problems including aggression and to make appropriate matches with potential adopters (Bennett, Litster, Weng, Walker, & Luescher, 2012; Bollen & Horowitz, 2008; Christensen, Scarlett, Campagna, & Houpt, 2007; Diederich & Giffroy, 2006; Dowling-Guyer, Marder & D’Arpino, 2011; Jones & Gosling, 2005; Rayment, Groef, Peters, & Marston, 2015; Taylor & Mills, 2006; Valsecchi, Barnard, Stefanini, & Normando, 2011). These evaluations are most often battery assessments with a number of subtests designed to test how the dogs would react to various types of stimuli they might encounter after adoption (Bennett et al., 2012; Dowling-Guyer, et al., 2011; Mornement, Toukhsati, Coleman, & Bennett, 2009). Behavior evaluations are sometimes referred to as temperament tests, implying that their results reveal behavioral traits that will be consistent over time (Bennett et al., 2012; Taylor & Mills, 2006). Other experts refer to them as behavior tests or evaluations that reflect changeable behaviors as opposed to consistent personality profiles (Bennett et al., 2012). Regardless, the animals are tested in experimental situations in which they are exposed to stimuli to elicit reactions, often in scenarios intended to mimic home environments or human interactions (Bennett et al., 2012; Diederich & Giffroy, 2006).

Research has indicated that behavior evaluations provide better information about dogs’ behavioral tendencies than the opinions of shelter staff (Van der Borg, Netto, & Planta, 1991) and are a good supplement to the dog’s behavioral history and demographic information (Bollen & Horowitz, 2008) in predicting aggression retrospectively (identifying aggression where there was a history of it) as well as reducing returns to the shelter for aggression and other reasons.
However, research has found poor prediction of post-adoption behavior (Bollen & Horowitz, 2008; Hennessey, et al., 2001; Mornement, Coleman, Toukhsati & Bennett, 2014) based on behavior assessed in the shelter. Additionally, researchers point out that no test will predict future behavior with absolute certainty (Bollen & Horowitz, 2008; Valsecchi et al., 2011).

One factor that affects the evaluations’ validity as predictors of post-adoption behavior is whether the tests are reliable measures in the first place. The reliability of a behavior evaluation is essential to its quality in measuring behavior (Diederich & Giffroy, 2006; Martin & Bateson, 1986; Taylor & Mills, 2006). Indeed, reliability is considered a prerequisite for validity (Jones & Gosling, 2005; Taylor & Mills, 2006). Reliability reflects the degree to which the test is free of measurement errors that can come from several points in the measurement process: the evaluators of the behavior, the subjects of the evaluation, or the evaluation itself. Intra-rater reliability has to do with the degree to which an individual observer is consistent within him or herself in repeatedly measuring the same behavior (Martin & Bateson, 1986; Taylor & Mills, 2006). Inter-rater reliability entails the degree to which multiple observers are consistent among each other in measuring behavior (Martin & Bateson, 1986; Taylor & Mills, 2006). Other aspects of reliability include measures of behavioral consistency within the dogs (test-retest) and consistency in the measures themselves (internal consistency). This study focuses on intra- and inter-rater reliability.

Researchers examining the efficacy of canine behavior evaluations commonly used in shelters generally have not reported the reliability of the evaluators applying the tests (Diesel, Brodbelt, & Pfeiffer, 2008; Jones & Gosling, 2005; Rayment et al., 2015; Taylor & Mills, 2006). In their 2005 review of research on canine personality and temperament, Jones & Gosling reported that much of the research that referred to inter-rater reliability did not include numerical
indices or contained incomplete evaluations of rater reliability and so could not be included in their analysis. Only three studies met their criteria for complete reporting. This lack of rater reliability analysis is not unusual in non-human animal research. Indeed, reporting of inter-rater reliability in animal behavior research was determined to be scant compared to human psychological research in a 2009 review of inter-rater reliability in studies published in two volumes of the academic journal *Animal Behaviour* (Kaufman & Rosenthal, 2009). Ninety-six of 100 articles reviewed did not report inter-rater reliability in their text (p. 1487). In a 2015 review of the state of canine behavior evaluations, Rayment et al., found that although behavior evaluations are frequently performed on dogs, few of the evaluations use accepted psychological reliability protocols that include consistent use of terminology and detailed assessment of relevant behaviors and contextual information.

Evaluators may vary in their observation and measurement of an animal’s behavior for many reasons. One issue is evaluators’ experience with the animals in question. One study related to observers of dog behavior indicated that level of experience with dogs did not significantly affect ability to correctly identify behavior, but that significant individual differences existed regardless of observer background (Tami & Gallagher, 2009, p. 159), which could presumably affect both intra- and inter-rater reliability. One recent study of inter-rater reliability of dog behavior found that non-experts could rate behavior reliably using standardized behavior evaluations (Fratkin et al., 2015). However, in a study of intra- and inter-rater reliability of dog behavior evaluations at a UK dog shelter, Diesel et al. (2008) found that inter-rater reliability was greater when staff had formal training in using the evaluation and had more than eight years of experience in dog training or behavior. In their review of canine behavior evaluations, Rayment et al. (2015) concluded that the crux of the issue of rater experience and
reliability may have to do with the type of behaviors being observed. Citing other research, Rayment et al. state that observers with more experience may be more accurate in identifying subtle behaviors and those that demonstrate abstract concepts such as “focus” than less experienced observers, implying that intra- and inter-rater reliability may be improved with increased evaluator experience. Variation across raters can occur for reasons other than experience, however. Veterinary students were found to have been biased by differing verbal contextual information in their scoring of identical behavior in behavioral tests of pigs, cows and chickens (Tuyttens et al., 2014). Students were given false or true information about such conditions as why animals were selected for breeding or the ambient temperature in the room when filming took place, and their behavior scores differed in keeping with the information provided. Factors such as observer gender in a study of salamanders (Marsh & Hanlon, 2004) and physical backgrounds used in videotapes of pig behavior (Wemelsfelder, Nevison, & Lawrence, 2009) have been noted to affect rater scores as well. And, dogs themselves may simply behave differently with evaluators of differing genders (Wells & Hepper, 1999) and familiarity (Kerepesi, Doka & Miklosi, 2015).

Some behavior evaluations for pet dogs in animal shelters have been assessed for validity either by comparing test results with owner questionnaires regarding their new dogs’ behavior (Christensen et al., 2007; Ledger & Baxter, 1997; Van der Borg, et al., 1991) or retrospectively with behavioral histories (Bollen & Horowitz, 2008). However, these and other studies, including the Hennessey et al. (2001) study mentioned earlier that found little agreement between shelter and in-home behavior, neglected to conduct or include information on rater reliability. In a 2006 review of the development of temperament tests for companion dogs, Taylor & Mills found fewer than ten studies of pet dog behavior evaluations in the peer-reviewed
literature (p. 104). Those they found reported moderate to good predictive validity in many tests, but the studies generally neglected the issue of inter- or intra-rater reliability. The authors point out that without reliability analyses “there is no information regarding the extent to which the validity of the test results has been affected by inconsistency on the part of the tester…” (p. 104). Indeed, a 2011 study of a temperament test for re-homed dogs claimed to be the first validation of a canine temperament test to address inter- and intra-rater agreement as well as test-retest reliability and validity (Valsecchi et al, 2011).

Where rater reliability has been assessed, results vary. In their 2005 meta-analysis, Jones & Gosling found that inter-rater reliability in canine personality research varied among studies and traits evaluated, with agreement ranging from poor to high. Based on the studies that reported on inter-rater reliability, they concluded that good reliability is achievable with canine behavior evaluations. These results indicate that while rater reliability on an evaluation may be good overall, there may be subtests in which rater reliability is low. The Valsecchi et al. study (2011) found a substantial average rater reliability overall but rater reliabilities of various behavioral categories evaluated ranged from fair to almost perfect, also indicating that some behaviors were less reliable for multiple raters to evaluate than others. That said, inter-rater reliability correlations in personality studies of a range of species, including humans, compare well with reliability correlations in canine behavior evaluations. In his 2001 review of animal personality studies, Gosling found moderate levels of inter-rater reliability among 21 studies and asserted that this “compares favorably” with inter-rater reliability correlations in human personality research (p. 60). Gosling noted, however, that quantitative evidence of intra-rater reliability correlations was still inadequate.

So, while the literature indicates better prediction of pet dog behavior using the results of
standardized behavior evaluations as well as other information such as behavior histories than without it, the validity of the evaluations may be improved further with attention to intra- and inter-rater reliability in assessment development. Rater reliability results may indicate subtests that weaken overall validity because raters are not reliable individually or when compared to each other. Hence, reliability analysis may lead assessment developers to drop weak subtests or alter them for improved rater reliability, thereby increasing the validity of the evaluation as a whole. This study evaluates the intra- and inter-rater reliability of the American Society for the Prevention of Cruelty to Animals’ (ASPCA) behavior evaluation for dogs exhibiting extreme fear. Fearful dogs enter animal shelters in high numbers, especially when hoarding situations, puppy mills, and dog fighting rings are investigated by law enforcement. Reliable and valid assessment of these dogs is important in deciding whether they should be placed for adoption and, if so, with what type of family composition, i.e., whether or not they should contain young children or other pets. Appropriate placement improves public safety, the dogs’ welfare, and adopters’ success with their new family member. This study represents a first look at the rater reliability of the ASPCA’s fearful dog evaluation. It is hypothesized that evaluators who have ongoing hands-on experience using the evaluation will display higher intra- and inter-rater reliability than those who do not. A second hypothesis is that there will be subtests for which inter-rater reliability is too low to support retaining the subtests in the evaluation as currently conducted.

**Methods**

The ASPCA’s behavior evaluation for fearful dogs is being developed and tested at its Behavior Rehabilitation Center in Madison, New Jersey, for use by shelters nationally for evaluation of fearful dogs. Most of the dogs have been rescued from hoarding, puppy mill and
dog fighting situations, although many have been referred to the center from other shelters and rescue groups. Once at the center, the dogs become part of a systematic rehabilitation program, the goal of which is to ready them for adoption. Dogs are assessed when they enter the program and then again every three weeks to track their progress until they either graduate from the program—by far the majority of them—or are euthanized for behavioral and/or health reasons.

The evaluation

Twenty-nine categories of behavior are rated with the Center’s evaluation tool, which was devised by senior ASPCA behavior staff, drawing somewhat from other behavior evaluations such as the ASPCA’s Safer Evaluation (www.aspcapro.org/safer-assessment-items), Sue Sternberg’s Assess-A-Pet evaluation (2006), and the Canine Behavioral Assessment & Research Questionnaire (www.guidedogs.org/wp/wp-content/uploads/c-barq-assessment.pdf), but also including unique scenarios. The evaluation is designed to test specific behavioral responses to stimuli in different contexts, which has been recommended by researchers (Rayment, et al., 2015) as preferable to evaluating behavior in single contexts. The fearful dog evaluation is administered via ten subtests conducted in three different settings. Dogs are evaluated in their kennels, rooms designated ‘unfamiliar rooms’ that contain minimal furnishings such as might be found in a home, and an outdoor chain link pen, providing for assessment of similar behaviors and situations in two indoor and one outdoor context. The kennels, rooms and pen are all marked with tape to designate near, medium and far distances that the dogs may move from evaluators and those distances are noted as part of the evaluation of the dogs’ levels of sociability (see Appendix A, beginning on page 45). Fear-based body postures and behaviors are assessed as part of the determination of a dog’s level of fear, anxiety or aggression (see Appendix A, pages 46
and 47). Dogs are also evaluated as they walk, or are sometimes carried in a crate if they are too fearful to walk, to and from those areas.

The ten evaluation subtests are summarized (with abbreviations in parentheses for single-context rating categories described further after this summary and in the appendices) as follows:

1. Person outside kennel (Out kennel): The evaluator approaches the kennel and stands outside of the kennel, looking at the dog. The evaluator then crouches and turns sideways at the kennel door and with palm up by her knee says “Here puppy, puppy, puppy” twice and is otherwise quiet. The evaluator may pet the dog briefly through the kennel if solicited. Otherwise, she stays quiet. Remaining crouched sideways, the evaluator repeats the call with treats in hand. If the dog takes a treat, a second can be offered, otherwise the evaluator sits quietly. Finally, the evaluator enters the kennel (Enter kennel), observing the dog’s reaction.

2. Person inside kennel (Inside kennel): After the evaluator has entered the kennel, she speaks quietly to the dog and crouches near the door turned to the side. With a palm up by her knee, she repeats the previous call and otherwise remains quiet. If the dog solicits attention, the evaluator may stroke it three times. Afterwards, she repeats the call with treats as when outside of the kennel. The dog is leashed (Leashing), walked out of the kennel if the dog is able (Leash indoors). If not, the dog is picked up or carried in a crate for this portion of the test.

3. Going outside: The dog is walked or carried through a door (Doorway) to an outdoor run area (Leash outdoors). The dog is walked or carried to the base of a set of portable stairs and the evaluator attempts to walk the dog up and down four stairs (Stairs). The dog is then walked or carried into an outdoor pen.
4. Outdoor pen 1: The evaluator and dog enter the pen, and the evaluator sits quietly in the center of the pen, ignoring the dog (Person outpen). If the dog solicits attention, the evaluator may stroke it three times while speaking softly and repeat that attention if solicited again. Otherwise, the evaluator remains still. Then, still sitting in the center of the pen, the evaluator calls “here, puppy, puppy, puppy” twice. If the dog solicits attention, the evaluator can stroke it three times and repeat that if the dog solicits again. Otherwise, the evaluator remains quiet. The evaluator then repeats the call with treats as in previous subtests.

5. Outdoor pen 2: Still sitting in the pen, the evaluator wiggles a squeaky toy on the ground, saying “Wanna play? Get it!” and squeaking the toy (Toy outdoor). This is attempted three times. If the dog tugs, the evaluator may play for ten seconds. The toy is then tossed away from the dog and allowed to get it if inclined. The evaluator then handles the dog gently, including stroking the dog’s cheek or neck, as well as his head, ears, back and tail (Outpen handling). The evaluator then strokes a back and front leg down to the feet and if the dog displays little to no fear of handling thus far, the evaluator pinches the web of a front foot. Finally, the dog is left alone in the pen for one minute (Alone outpen).

6. Unfamiliar room 1: The dog is walked or carried into a furnished, unfamiliar room. The evaluator sits quietly in the corner furthest from the door, ignoring the dog (Person unfamiliar). If the dog solicits attention, the evaluator gives him three strokes while speaking softly. This is repeated if the dog solicits again. Otherwise, the evaluator remains still. The evaluator calls to the dog without treats and then calls with treats as in previous subtests.
7. Unfamiliar room 2: Still sitting, the evaluator invites the dog to play with a squeaky toy as in the outdoor pen, holding it first and then tossing it (Toy indoor). The evaluator again gently handles the dog, moving through the same stroking process as previously described (Unfamiliar handling). The dog is then left alone in the unfamiliar room for one minute (Alone unfamiliar).

8. Novel person greeting: After the evaluator and scribe return to the room, a person unknown to the dog knocks on the door and enters (Novel greeting). The evaluator greets the novel person and the two sit on a couch in the room. If safe, the tester drops the leash. With hand held out and palm up, the novel person calls to the dog in the same manner as previously and follows that up with calling to the dog with treats in hand. If the dog does not take the treat, the unknown person places the treat on the floor. The novel person then leaves the room.

9. Familiar person greeting: The greeting test is repeated with a person familiar to the dog (Familiar greeting). The dog is returned to its kennel.

10. People standing at kennel and novel object: With the dog in the indoor portion of the kennel, the dog is left alone for a minute to settle. The novel person approaches and stands in front of the kennel, looking at the dog (Novel kennel). The person repeats the call, “Here puppy, puppy, puppy” two times and then walks away. Following that, the familiar person approaches the kennel and repeats this scenario (Familiar kennel). The door separating the indoor and outdoor portions of the run is then opened, revealing a novel object and food bowl containing cat food in the doorway. The tester walks out of sight and the dog’s response to the object is filmed for one minute (Novel object).
From these subtests, 21 single-context numeric ratings are given based on the dog’s response to specific stimuli. Data points related to fear, aggression and sociability are entered for the dog’s response in each of the single contexts and, from those, the dog is rated on a 1-3 scale, with 1 indicating ‘not concerning’, 2 designated a ‘red flag’ (concerning), and 3 labeled ‘alert’ (very concerning) (see Appendix B, page 48). The single contexts are indicated by the abbreviations above in parentheses. At the end of the evaluation, seven general categories of behavior, such as handling in both contexts in which it is assessed, are graded with a letter grade of A, B, C, or D, and the dog is also assigned an overall behavior grade for the entire evaluation (see Appendix B, page 49). For the letter grades, A indicates no concerning behavior in that category, B means mild fear, anxiety or aggression, C connotes moderate fear, anxiety and aggression, and D indicates severe fear, anxiety or aggression. Behavioral definitions for each rating, grade, and behavior evaluated are further specified on the evaluation for standardization purposes (see Appendix B, beginning on page 48). To see visual representations of what some of that data look like as rater comparisons, see Figures 1A, 1B, 2A and 2B, pages 39 and 40.

Hence, three gradations of behavior are evaluated. For example, how the dog reacts to handling is evaluated in two contexts (in the outdoor pen and in an unfamiliar room—“Outpen handling” and “Unfamiliar handling,” respectively), rated individually in each of those, and then a dual-context handling grade—“Handling unfamiliar”—is assigned. A grade for socialization to familiar people—“Socialization familiar”—is assigned from the ratings for “Familiar greeting” and “Familiar kennel.” A grade for socialization to unfamiliar people—“Socialization unfamiliar”—is derived from the eleven ratings in the contexts related to interactions with the evaluator, in the unfamiliar room and outside pen, and greetings by novel people. A grade for leash walking—“Leashwalk unfamiliar”—is based on ratings from the “Leash indoors,” “Leash
outdoors,” “Doorway,” and “Stairs” contexts. A “Leashing unfamiliar” grade is based on the “Leashing” context rating. A grade for the dog’s reaction to familiar people at the kennel—“Kennel presence familiar” is based on the rating for the “Familiar kennel” context. A grade for the dog’s reaction to unfamiliar people at the kennel—“Kennel presence unfamiliar”—stems from ratings for the “Out kennel,” “Enter kennel,” and “Novel kennel” contexts. A handling grade, “Handling unfamiliar,” is given based on ratings for “Outpen handling” and “Unfamiliar handling.” The overall grade is based on the grades in all contexts as well as a determination of the dogs’ fear of novelty, outdoors and indoors through the relevant behavior contexts (see Appendix C, page 50 and 51).

An evaluator and a scribe conduct the evaluations, and the scribe or a third person handles a video recorder. All evaluations are recorded and the videos are stored on a master hard drive. Video recorded evaluations of 12 dogs of various breeds, sexes and ages were selected for this study. Criteria for the selected recordings were that they represented a range of behaviors according to grades that the dogs were assigned at the time of the evaluations and that they were of good quality so that dogs’ behavior could be seen clearly. Additionally, the tests had to fall within a seven-month period during which the process was no longer new to the staff and hence ran similarly each time, but also so that the dogs themselves had been out of the program for some time to reduce bias based on memory of the dogs or their evaluations. The center opened its doors in March 2013, and the selected tests occurred between December 2013 and July 2014. The recordings, which were stored on a dedicated hard drive, were edited for this study using Cyberlink Power Director 13 software. We removed any identifying information such as the dog’s name and evaluation number and added titles for each subtest so the raters could clearly determine when each subtest began and ended.
Participants

Four raters viewed and rated the behavior of the 12 dogs on the 12 selected video recordings. All raters were familiar with the test and highly experienced in working with dogs. All were adult women. Two raters regularly conducted the Rehabilitation Center evaluations and two others were senior behavior personnel with the ASPCA who participated in developing the evaluation but did not test the Rehabilitation Center dogs at the time of this study. The two raters who regularly conducted the evaluations had different degrees of experience: one had approximately 25 years of experience assessing dog behavior and the other two years, although she also had about 10 years of experience as a dog trainer. The two raters who did not conduct evaluations had ten and eleven years of experience with canine behavior evaluations.

Data collection

The raters were given instructions on accessing, watching and rating the recorded behavior, including that the videos should be watched with the sound off so as to eliminate hearing comments between the assessor and scribe that would bias study results. This meant, however, that they also could not hear vocalizations from the dogs. Additionally, the raters were instructed not to revise an evaluation once they had completed it, but that reviewing video segments during the rating process was acceptable. Raters accessed the videos through a password-protected proprietary online, shared folder and each was assigned a randomized order for viewing the recordings. Viewing order for each rater was determined by pulling video numbers out of a hat, one at a time, until none remained. This process was repeated for each rater. They watched the videos and rated the dogs’ behavior on two separate occasions four to five weeks apart between August 2015 and January 2016. Each rater entered her data into a separate database file that was stored and accessed in the same password-protected, shared online
folder as the video recordings. Data were downloaded from each database file and compiled in a Microsoft Excel spreadsheet for analysis. The numeric ratings (1, 2 or 3) from the 21 evaluation contexts and the alphabetic grades (A, B, C or D transformed to 1, 2, 3 or 4 for analysis) for the seven categories of behavior as well as the dogs’ overall grades were analyzed for this study.

**Data analysis**

Data were analyzed using agreement, unweighted Kappa, quadratic weighted Kappa and Pearson \( r \) correlation measures. Researchers recommend using indices of both agreement (the extent to which raters make the same judgment) and reliability (the degree to which raters are similar in their judgments) in analyzing and reporting intra-rater and inter-rater reliability (Tinsley & Weiss, 1975; Kaufman & Rosenthal, 2009). Agreement measures, which are percentages of occurrence of the same paired judgments, can inflate agreement because they do not take into account the statistical probability of agreement by chance alone or the relative chance for agreement based on a rating’s placement on a scale. Unweighted Kappa measures the proportion of agreement after removing the proportion of agreement that would be expected by chance. Quadratic weighted Kappa goes further than that, taking into account that use of middle ratings of a scale may be more subject to error than those on the extremes of a scale (Banerjee, Capozzoli, McSweeney & Sinha, 1999; Cohen, 1968). As such, quadratic weighted Kappa has been recommended for ordinal data, such as used in this study, where there is an element of subjectivity to the scales because of rater perception of their meaning and thus more potential for error in middle ratings. Finally, the commonly used Pearson \( r \), which is similar to quadratic weighted Kappa but not designed for use with ordinal data, has been a subject of recent debate as a measure of correlation in research generally (Puth, Neuhauser & Ruxton, 2014). Hence, this study includes both the quadratic weighted Kappa and Pearson \( r \) for comparative purposes;
unweighted Kappa is included to illustrate patterns in statistical analysis. Kappa measures tend to fall on a scale of zero to one with zero indicating agreement no better than chance and one indicating perfect agreement (Chodorow, personal communication). There can be negative Kappa statistics if agreement is worse than chance, but that occurs rarely. Pearson $r$ correlations are on a scale of negative one to one, with negative one indicating perfect negative agreement in which one value in a correlation rises as the other falls, zero meaning no agreement, and one indicating perfect positive agreement in which the values rise equally (Field, 2009).

**Results**

Agreement, unweighted Kappa, quadratic weighted Kappa and Pearson $r$ statistics were calculated for each of the 29 behavioral categories previously mentioned. The results for each measure were organized into a $29 \times 4$ (rater) matrix for testing the statistical significance of intra-rater reliability and into a $29 \times 6$ (rater pairs) matrix for testing the significance of inter-rater reliability. The 29 behavioral categories were treated as the replication variable and the raters or rater pairs were treated as repeated measures variables.

Both intra-rater and inter-rater agreement and reliability statistics were generally good to excellent across the raters and rater pairs. Intra-rater agreement for the four raters ranged from 0.72 to 0.95. Friedman tests indicated that there were significant differences among the four raters ($\chi^2 (3) = 47.04, p < .001$) on agreement, on quadratic weighted Kappa, which ranged from 0.71 to 0.96 ($\chi^2 (3) = 49.16, p < .001$), and on Pearson $r$, which ranged from 0.79 to 0.96 ($\chi^2 (3) = 40.84, p < .001$) (see Table 1 for intra-rater reliability means, page 33). The unweighted Kappa values, which ranged from 0.51 to 0.97, were also significantly different ($\chi^2 (3) = 41.09, p < .001$), but showed a greater range and generally lower values than the quadratic weighted Kappa, reflecting the statistic’s limitation in evaluating agreement differences in relation to chance as
opposed to taking into consideration the possibility for greater and lesser disagreements based on extremes and similarities in scale rankings. According to one accepted interpretation of Kappa values, a range of 0.00 to 0.20 represents slight agreement, 0.21 to 0.40 signifies fair agreement, 0.41 to 0.60 reflects moderate agreement, 0.61 to 0.80 represents substantial agreement, and 0.81 to 1.00 constitutes almost perfect agreement for Kappa measurements (Landis & Koch, 1977). Another interpretation puts values of 0.75 and greater as representing excellent agreement, values of 0.40 – 0.75 as signifying fair to good agreement, and values lower than 0.40 as reflecting poor agreement (Banjeree et al., 1999).

As noted above, for the intra-rater measures, the Friedman test analyses of the differences among raters were statistically significant at $p < .001$ for all measures, confirming that the raters were not all equal in their intra-rater reliability. As a follow up to this, Wilcoxon signed ranks tests were used to compare each rater to each of the other raters. Differences between the two current use raters were not significant on any measure, but differences between all other rater combinations were significant on all measures at $p < .05$ or better. Rater 3 had significantly higher intra-rater reliability than the other three raters, and rater four had significantly lower intra-rater reliability than the other three raters. The differences between the means of the current use raters, raters one and two, and the non-current use raters, raters three and four, were not significant for any measure.

Inter-rater agreement and reliability ranged similarly to intra-rater values among the six rater pairs (see Tables 2 and 3 for inter-rater reliability means, page 34). Agreement measures for all six rater pairs for the first evaluation of recorded behaviors ranged from 0.68 to 0.82, and were significantly different by a Friedman test ($\chi^2 (5) = 36.99, p < 0.001$). For the second evaluation, they ranged from 0.67 to 0.79 and were also significantly different ($\chi^2 (5) = 18.158, p$
Quadratic weighted kappa values ranged from 0.64 to 0.85 ($\chi^2 (5) = 38.809, p < 0.001$) for evaluation one, and 0.63 to 0.71 ($\chi^2 (5) = 21.646, p = 0.001$) for evaluation two. Pearson $r$ correlations again were very similar to quadratic weighted Kappa with a range of 0.67 to 0.87 ($\chi^2 (5) = 37.153, p < 0.001$) for evaluation one and 0.72 to 0.83 ($\chi^2 (5) = 17.803, p < 0.01$) for evaluation two. As with the intra-rater reliability numbers, unweighted Kappa numbers were generally lower than quadratic weighted Kappa and Pearson $r$, with a range of 0.45 to 0.69 ($\chi^2 (5) = 35.936, p < 0.001$) for the first evaluation and 0.43 to 0.65 ($\chi^2 (5) = 19.085, p < 0.01$) for the second evaluation. All sample sizes were 29, the total number of behavioral categories rated.

The Friedman test analysis of the differences among mean ranks of the rater pairs was statistically significant at $p < 0.001$ for all levels of analysis in the first evaluation, confirming that the pairs were not equal in their inter-rater reliability. In the second evaluation, quadratic weighted Kappa means were significant at $p < 0.001$, while the agreement, unweighted Kappa and Pearson $r$ correlation means were significant at $p < 0.01$. Finally, in all pairwise rankings in this analysis, the current use pair, raters one and two, ranked highest although the non-current use pair, raters three and four, were not necessarily lowest.

Planned comparisons of rater pairs using the Wilcoxon signed rank test confirmed significant differences among rater pairs. For the agreement analysis, the difference between the current use pair and the non-current use pair for the first evaluation was significant at $Z = 4.10, p < 0.001$ and between the current use pair and the mean of the other five rater pairs at $Z = 4.57, p < 0.001$. In the second evaluation, agreement differences remained significant, with the difference between the current use pair and the non-current use pair at $Z = 2.15, p < 0.05$ and the difference between the current use pair and the mean of the other five pairs at $Z = 3.41, p = .001$. 
For the quadratic weighted Kappa statistics, the differences between the current use pair and the non-current use pair were again significant ($Z = 4.42, p < 0.001$), as were the differences between the current use pair and the mean of all other pairs ($Z = 4.70, p < 0.001$), in the first evaluation. Differences remained significant in the second evaluation: $Z = 2.83, p < 0.01$ for the current user versus non-current user pairs and $Z = 3.99, p < 0.001$ for the current use pair and the mean of all others. Results were similar for the Pearson $r$ correlations. For the first evaluation, the differences were significant between both the current rater pair and non-current rater pairs ($Z = 3.89$) and the current pair versus the mean of the other pairs ($Z = 4.21$) at $p < 0.001$. Results were still significant, although less so, in the second evaluation: for current use raters versus non-current use raters ($Z = 2.20, p < 0.05$) and current use raters compared with the mean of all others ($Z = 3.23, p = 0.001$). Unweighted Kappa comparisons were also significant. For the first evaluation, current versus non-current rater pairs were significant at $Z = 4.57, p < .001$ and current versus the mean of all other rater pairs were significant at $Z = 4.68, p < .001$.

Significance held for the second evaluation for current versus non-current raters ($Z = 2.20, p < .05$) and current versus the mean of all other rater pairs ($Z = 3.49, p < 0.001$).

The same four statistical tests were conducted for the behavioral categories rated and analyzed for this study. All are reported in tables for inter-rater means by behavior categories and evaluation one or two, and by behavior categories and current use versus non-current use rater pairs (see Tables 4 and 5, pages 35 and 37). However, in determining which categories have low reliability to support retaining them as is in the behavioral test, we will focus on the quadratic weighted Kappa statistics, which we have already seen are similar to the Pearson $r$ correlations (see Figures 3A, 3B, 4A and 4B, beginning on page 41). Five subtests have unacceptably low quadratic weighted Kappa results, that is, below 0.60, above which represents substantial
agreement according to Landis & Koch (1977). The Alone Outpen subtest (when the dog is left alone in the outdoor pen) has quadratic weighted Kappa means of 0.32 and 0.25 for evaluations one and two, respectively. The Alone Unfamiliar subtest (when the dog is left alone in the unfamiliar room) has quadratic weighted Kappa means of 0.40 on evaluation one and 0.23 for evaluation two. The Novel Object subtest (when the dog encounters a novel object) was similarly low, with a quadratic weighted Kappa of 0.23 and 0.37 for evaluation one and two, respectively. Subtests involving the toy were better but still represent just moderate (Landis & Koch, 1977) or fair to good agreement (Banjeree, et al., 1999), depending on the interpretation one uses. For Toy Indoor, quadratic weighted Kappa means were 0.51 and 0.52, and for Toy Outdoor the means were 0.58 and 0.53, for evaluations one and two, respectively.

In general, the quadratic weighted Kappa values for particular subtests were consistent with rates of agreement (see Table 4, page 35), that is, where one was high, so was the other and vice versa. However, agreement and quadratic weighted Kappa numbers for some subtests were mismatched in ways that should be considered. For example, in the Alone Unfamiliar subtest, there is an agreement value of 0.81 that corresponds with the quadratic weighted Kappa of 0.40 for the first evaluation, likely indicating few mismatched ratings (so high agreement), but the mismatches that exist differ widely, such as ratings of one versus three rather than one versus two, (so low quadratic weighted Kappa) (Chodorow, personal communication). On the other hand, Leashwalk Unfamiliar has a quadratic weighted Kappa mean for both evaluations of 0.76 that corresponds with a low agreement mean of 0.59, presumably related to the frequency of adjacent mismatches in ratings (one versus two and two versus three). And the Kennel Presence Unfamiliar category has a barely acceptable quadratic weighted Kappa of 0.60 for the second evaluation that corresponds with a low agreement value of 0.51, indicating both low levels of
matching ratings as well as large mismatches in ratings.

Looking at the means by behaviors and current versus non-current use rater pairs (Table 5, page 37), some but not all of these subtest results remain low across both rater pairs, but the non-current rater pair adds some additional subtests to the tally of unacceptably low quadratic weighted Kappa means. The Alone Outpen subtest had a quadratic weighted Kappa of 0.50 for the current use pair and 0.00 for the non-current use pair, indicating no greater agreement than chance. Alone Unfamiliar had a quadratic weighted Kappa of 0.56 for the current use pair but a 0.26 for the non-current user pair. No other low quadratic weighted Kappa means are reported for any of the behavioral subtests for the current use pair. However, four other categories have low quadratic weighted Kappa means from the non-current use rater pair: Doorway, in which the dog is encouraged to walk through a doorway, (0.42); Leash Indoors, in which the dogs walked on leash indoors, (0.45); Toy Outdoor (0.50) and Leashwalk Unfamiliar, which was the grade given to the dog’s ability to walk on leash with the evaluator who is supposed to be unfamiliar to the dog, (0.54). These results support previous analyses of differences among the rater pairs in relation to their currency in using the Rehabilitation Center’s behavioral evaluation.

Discussion

In general, the results support the study’s hypotheses, that the evaluators with ongoing hands-on experience using the behavior evaluation would have higher intra- and inter-rater reliability than those who do not use the evaluation regularly, and that some subtests in the evaluation would have inter-rater reliability that is too low to retain them in the evaluation as is. Although one rater not currently using the evaluation had the highest intra-rater reliability measurements (rater 3, with a quadratic weighted Kappa of 0.96 compared with 0.83, 0.86 and 0.71 for the others), inter-rater reliabilities for the rater pair who currently use the evaluation on a
regular basis are consistently significantly higher than for all other rater pairs, who include the non-current use pair and combinations of current and non-current use raters. Hence, the rater with highest reliability internally is not in reliable agreement with the other raters. Among the raters using this behavior evaluation, those who use it regularly and often have the greatest inter-rater reliability, even though all raters have years of behavioral experience with dogs and one of the current-use raters, in fact, has the fewest years of experience of the four.

There are a variety of possibilities for the low inter-rater reliabilities on some subtests. According to previous research, variation in reliability in rating some behaviors may have to do with the nature of the behaviors being observed (Rayment et al., 2015), with subtle behaviors being less reliably evaluated among observers. A corollary is that the more “subtle” behaviors may not be well defined in the evaluation. In this study, raters had low inter-rater reliabilities in the Alone Outpen and Alone Unfamiliar subtests, in which dogs’ reaction to being left alone in an outdoor pen or unfamiliar room were rated as either not concerning (essentially, behavior expected of a normal dog), red flag (mild issues adoption counselors would like to tell adopters about), or alert (severe issues they would need to tell adopters about). While there are some further behavioral definitions given (see Appendix B, page 48), it may also be that what normal behavior as opposed to mild issues or mild issues as opposed to severe ones look like for a dog in these contexts is a subtle distinction not easily observed or not adequately defined by the evaluation for the situation. The Novel Object and two toy tests also had low inter-rater reliabilities. Again, what constitutes normal behavior in these situations? Ignoring the items could be considered normal, but would interacting with them be better? That is not clear in the not concerning, red flag, and alert definitions. Similar factors may be at play where inter-rater reliabilities were low for the non-current use raters but acceptable for the current use raters, such
as the Doorway and Leash Indoors subtests (see Table 5, Inter-rater means by behavior categories, and current evaluator and non-current evaluator pairs, page 37, and Figures 3A and 3B, pages 41 and 42). Is it “normal” for a dog to simply walk through the doorway, or should the dog move through it in a particular way? What is a mild versus severe problem for a dog moving through a doorway? And, what particular behaviors are relevant to those categories for leash walking? These may all need to be defined more specifically in the ratings definitions. The only multi-context category that was graded that had a low inter-rater reliability was Leashwalk Unfamiliar, and that may need to be evaluated similarly. Are the grading definitions precise enough to achieve good reliability for this behavioral category?

In considering which subtests to drop or alter, and how to alter subtests, mismatches in agreement and quadratic weighted Kappa values should be taken into consideration. Both wide disagreement among raters (low quadratic weighted Kappas) and frequent disagreement (low agreement measures) may give important indications for changing evaluation criteria to improve rater reliability and evaluation validity. It is also important to consider the relative importance of each subtest. It may be more important to address problems related to rating dogs’ reactions to being left alone because of welfare implications for the dog and potential difficulties for adopters than it is to address poor levels of rater agreement related to a novel object test.

More broadly, the results of this study agree with those of previous research that have found moderate to high levels of inter-rater reliability (DePalma et al., 2005; Jones & Gosling, 2005; Kroll, Houpt & Erb, 2004; Mornement et al., 2009; Sinn, Gosling & Hilliard, 2010) and high levels of intra-observer reliability (Diesel et al, 2008;) in relation to canine behavior evaluations. It also supports previous studies that have found differing levels of agreement among and within observers in relation to individual subtests of an evaluation, (Fratkin et al,
2015; Sinn et al, 2010; Valsecchi et al, 2011). Reporting of intra-rater reliability in this study adds to the scant data related to this category of observer reliability in the literature. The results also illustrate that while intra-rater reliability is high enough to indicate that individual observers rate behavior consistently, inter-rater reliability may well be lower than intra-rater reliability. Hence, both should be investigated in relation to any particular evaluation.

While previous research has examined rater variability in the use of canine behavior evaluations in terms of evaluator backgrounds (e.g., owners, shelter employees, those who work with dogs) or years of experience working with dogs (Diesel et al., 2008; Fratkin et al., 2015; Tami & Gallagher, 2009), our hypotheses identify amount of use of the tool as a potentially important source of variability and our results appear to bear that out as a significant factor for our four raters. However, there are other factors possibly contributing to that variability. It is possible that some raters re-watched portions of the videos more than others before entering their data. In private correspondences after the study was conducted, rater one, who regularly tests the dogs and is the least experienced of the raters in terms of years spent evaluating dogs, estimated that she re-watched portions of videos between a fifth and quarter of the time. Rater two, by far the most experienced in terms of years of experience evaluating dogs and a regular evaluator, said she “rarely” re-watched video. These raters had quadratic weighted Kappa intra-rater reliabilities of 0.83 and 0.86, respectively, and together had a quadratic weighted Kappa inter-rater reliability mean of 0.83 for both evaluations. Raters three and four, whose experience evaluating dogs is comparable and neither of whom evaluated Rehabilitation Center dogs at the time of the study, re-watched videos in different amounts. Rater three, who had the highest intra-rater reliability, estimated she re-watched portions of video about 35 percent of the time, while rater four said she did not re-watch video “very often.” These raters had respective quadratic
weighted Kappa intra-rater reliabilities of 0.96 and 0.71, and the two together had an overall quadratic weighted Kappa inter-rater reliability score of 0.68. Raters one, two and four reported re-watching video segments only if they thought they had missed something or needed more information. Rater three, who seems to have re-watched more video segments than the others, said she did so when trying to decide on an appropriate fear level and when a subtest began and ended. Possibly, re-watching video segments more frequently and with these concerns in mind contributed to her high intra-rater reliability. Another factor in the rater variability may be that three of the four raters (one, two and three) work at the Rehabilitation Center and, even if they did not evaluate these dogs, did know them. Although the study design attempted to minimize the potential for remembering the dogs or their evaluations, rater four, who had the lowest intra-rater reliability score of the four raters, had had no contact with any of the dogs whose behavior was reviewed and rated.

To the extent that variability of rater reliability was affected by the amount of use of the evaluation as hypothesized, re-watching of video, and memory of the dogs, raters who are not practiced at using the evaluation and also do not know the dogs may be expected to have lower intra- and inter-rater reliability. One potential solution is that staff who rate dogs using this evaluation but do so infrequently could practice using the evaluation instrument with previously videotaped behavior to maintain familiarity with the evaluation itself. When doing live evaluations, agreement may be improved by watching video directly following the evaluation and with attention to detail about such important factors as fear, anxiety or aggression levels. If experience with the dogs is affecting reliability numbers, raters must work diligently to be objective in rating behavior. Finally, rater reliability in using the evaluations should be reexamined periodically to be certain criteria related to particular subtests are specific enough for
the raters to achieve high levels of reliability, particularly when new staff are hired as evaluators.

**Conclusion**

Intra- and inter-rater reliability of four raters using the ASPCA’s behavior evaluation of fearful dogs was generally quite good and in keeping with past research that has indicated strong intra- and inter-rater reliability of canine behavior evaluations. Results also agreed with past research in that they indicated that some specific subtests had unacceptably low rater reliability. While rater reliabilities tended to be highest for raters who frequently used the evaluation regardless of years of experience raters had conducting behavior evaluations, factors such as re-watching video and remembering the dogs may also have been at play. Additionally, it is possible that evaluation criteria were not specific enough for some behaviors for the raters to achieve adequately high levels of reliability on the subtests evaluating them.
References


Chodorow, Martin. (2016, April 14). Personal communication.


www.aspcapro.org/safer-assessment-items

Table 1

*Intra-rater Reliability Agreement Means Across All Dogs and All Behavior Categories for Each Rater*

<table>
<thead>
<tr>
<th></th>
<th>Current use rater 1</th>
<th>Current use rater 2</th>
<th>Non-current use rater 3</th>
<th>Non-current use rater 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement***</td>
<td>0.82</td>
<td>0.82</td>
<td>0.95</td>
<td>0.72</td>
</tr>
<tr>
<td>Unweighted Kappa***</td>
<td>0.68</td>
<td>0.71</td>
<td>0.92</td>
<td>0.51</td>
</tr>
<tr>
<td>Quadratic Weighted Kappa***</td>
<td>0.83</td>
<td>0.86</td>
<td>0.96</td>
<td>0.71</td>
</tr>
<tr>
<td>Pearson r***</td>
<td>0.86</td>
<td>0.88</td>
<td>0.97</td>
<td>0.79</td>
</tr>
</tbody>
</table>

*** n = 29, all means significant at p < 0.001
Table 2

**Pairwise Inter-rater Reliability Agreement Means Across All Dogs and All Behavior Categories for Each Rater Pair on Evaluation One.**

<table>
<thead>
<tr>
<th></th>
<th>Current use pair (raters 1 and 2)</th>
<th>Non-current use pair (raters 3 and 4)</th>
<th>Pair 3 (raters 1 and 3)</th>
<th>Pair 4 (raters 1 and 4)</th>
<th>Pair 5 (raters 2 and 3)</th>
<th>Pair 6 (raters 2 and 4)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement***</td>
<td>0.82</td>
<td>0.70</td>
<td>0.68</td>
<td>0.69</td>
<td>0.73</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>Unweighted Kappa***</td>
<td>0.69</td>
<td>0.47</td>
<td>0.46</td>
<td>0.45</td>
<td>0.54</td>
<td>0.54</td>
<td>0.53</td>
</tr>
<tr>
<td>Quadratic Weighted Kappa***</td>
<td>0.85</td>
<td>0.68</td>
<td>0.64</td>
<td>0.64</td>
<td>0.72</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Pearson r***</td>
<td>0.87</td>
<td>0.74</td>
<td>0.67</td>
<td>0.72</td>
<td>0.76</td>
<td>0.77</td>
<td>0.76</td>
</tr>
</tbody>
</table>

*** n = 29, all means significant at p < 0.001

Table 3

**Pairwise Inter-rater Reliability Agreement Means Across All Dogs and All Behavior Categories for Each Rater Pair on Evaluation Two**

<table>
<thead>
<tr>
<th></th>
<th>Current use pair (raters 1 and 2)</th>
<th>Non-current use pair (raters 3 and 4)</th>
<th>Pair 3 (raters 1 and 3)</th>
<th>Pair 4 (raters 1 and 4)</th>
<th>Pair 5 (raters 2 and 3)</th>
<th>Pair 6 (raters 2 and 4)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement**</td>
<td>0.79</td>
<td>0.72</td>
<td>0.71</td>
<td>0.67</td>
<td>0.71</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>Unweighted Kappa**</td>
<td>0.65</td>
<td>0.52</td>
<td>0.51</td>
<td>0.43</td>
<td>0.54</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>Quadratic Weighted Kappa**</td>
<td>0.81</td>
<td>0.69</td>
<td>0.71</td>
<td>0.63</td>
<td>0.70</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>Pearson r**</td>
<td>0.83</td>
<td>0.77</td>
<td>0.74</td>
<td>0.72</td>
<td>0.74</td>
<td>0.75</td>
<td>0.76</td>
</tr>
</tbody>
</table>

*** n = 29, all means significant at p = 0.001
** n = 29, all means significant at p < 0.01
### Table 4

*Inter-rater Means by Behavior Categories and Evaluation*

<table>
<thead>
<tr>
<th></th>
<th>Agr. 1</th>
<th>Agr. 2</th>
<th>K 1</th>
<th>K 2</th>
<th>K_w 1</th>
<th>K_w 2</th>
<th>r 1</th>
<th>r 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiar greeting</td>
<td>0.78</td>
<td>0.71</td>
<td>0.61</td>
<td>0.53</td>
<td>0.76</td>
<td>0.79</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>Familiar kennel</td>
<td>0.83</td>
<td>0.88</td>
<td>0.69</td>
<td>0.77</td>
<td>0.77</td>
<td>0.84</td>
<td>0.79</td>
<td>0.89</td>
</tr>
<tr>
<td>Alone outpen</td>
<td>0.72</td>
<td>0.72</td>
<td>0.18</td>
<td>0.14</td>
<td><strong>0.32</strong></td>
<td><strong>0.25</strong></td>
<td>0.90</td>
<td>0.79</td>
</tr>
<tr>
<td>Alone unfamiliar</td>
<td>0.81</td>
<td>0.63</td>
<td>0.23</td>
<td>0.13</td>
<td><strong>0.40</strong></td>
<td><strong>0.23</strong></td>
<td>0.41</td>
<td>0.26</td>
</tr>
<tr>
<td>Doorway</td>
<td>0.68</td>
<td>0.65</td>
<td>0.44</td>
<td>0.36</td>
<td>0.63</td>
<td><strong>0.54</strong></td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>Enter kennel</td>
<td>0.65</td>
<td>0.61</td>
<td>0.44</td>
<td>0.43</td>
<td>0.62</td>
<td>0.72</td>
<td>0.66</td>
<td>0.79</td>
</tr>
<tr>
<td>Inside kennel</td>
<td>0.75</td>
<td>0.72</td>
<td>0.61</td>
<td>0.59</td>
<td>0.77</td>
<td>0.80</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td>Leash indoors</td>
<td>0.63</td>
<td>0.67</td>
<td>0.33</td>
<td>0.48</td>
<td>0.61</td>
<td>0.66</td>
<td>0.62</td>
<td>0.70</td>
</tr>
<tr>
<td>Leash outdoors</td>
<td>0.76</td>
<td>0.78</td>
<td>0.57</td>
<td>0.66</td>
<td>0.80</td>
<td>0.83</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td>Leashing</td>
<td>0.79</td>
<td>0.83</td>
<td>0.64</td>
<td>0.75</td>
<td>0.81</td>
<td>0.89</td>
<td>0.82</td>
<td>0.92</td>
</tr>
<tr>
<td>Novel object</td>
<td>0.63</td>
<td>0.60</td>
<td>0.25</td>
<td>0.23</td>
<td><strong>0.23</strong></td>
<td><strong>0.37</strong></td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>Novel greeting</td>
<td>0.63</td>
<td>0.82</td>
<td>0.44</td>
<td>0.73</td>
<td>0.69</td>
<td>0.85</td>
<td>0.72</td>
<td>0.88</td>
</tr>
<tr>
<td>Novel kennel</td>
<td>0.72</td>
<td>0.74</td>
<td>0.57</td>
<td>0.58</td>
<td>0.74</td>
<td>0.74</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Out kennel</td>
<td>0.82</td>
<td>0.82</td>
<td>0.73</td>
<td>0.73</td>
<td>0.87</td>
<td>0.87</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>Outpen handling</td>
<td>0.76</td>
<td>0.64</td>
<td>0.61</td>
<td>0.45</td>
<td>0.85</td>
<td>0.70</td>
<td>0.87</td>
<td>0.72</td>
</tr>
<tr>
<td>Person outpen</td>
<td>0.72</td>
<td>0.79</td>
<td>0.55</td>
<td>0.63</td>
<td>0.73</td>
<td>0.75</td>
<td>0.79</td>
<td>0.77</td>
</tr>
<tr>
<td>Person unfamiliar</td>
<td>0.75</td>
<td>0.72</td>
<td>0.57</td>
<td>0.55</td>
<td>0.71</td>
<td>0.68</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>Person unfamiliar room</td>
<td>0.75</td>
<td>0.72</td>
<td>0.57</td>
<td>0.55</td>
<td>0.71</td>
<td>0.68</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>Stairs</td>
<td>0.89</td>
<td>0.92</td>
<td>0.81</td>
<td>0.86</td>
<td>0.90</td>
<td>0.93</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>Toy indoor</td>
<td>0.76</td>
<td>0.83</td>
<td>0.55</td>
<td>0.69</td>
<td><strong>0.52</strong></td>
<td><strong>0.51</strong></td>
<td>0.59</td>
<td>0.57</td>
</tr>
<tr>
<td>Toy outdoor</td>
<td>0.70</td>
<td>0.68</td>
<td>0.46</td>
<td>0.42</td>
<td><strong>0.58</strong></td>
<td><strong>0.53</strong></td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td>Unfamiliar handling</td>
<td>0.83</td>
<td>0.75</td>
<td>0.65</td>
<td>0.52</td>
<td>0.76</td>
<td>0.68</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>Socialization familiar</td>
<td>0.78</td>
<td>0.69</td>
<td>0.66</td>
<td>0.55</td>
<td>0.83</td>
<td>0.83</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td>Socialization unfamiliar</td>
<td>0.60</td>
<td>0.60</td>
<td>0.45</td>
<td>0.43</td>
<td>0.79</td>
<td>0.76</td>
<td>0.81</td>
<td>0.79</td>
</tr>
<tr>
<td>Leashwalk unfamiliar</td>
<td>0.58</td>
<td>0.60</td>
<td>0.40</td>
<td>0.46</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td>Leashing unfamiliar</td>
<td>0.71</td>
<td>0.67</td>
<td>0.55</td>
<td>0.50</td>
<td>0.88</td>
<td>0.85</td>
<td>0.91</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Note: Agr. = agreement measures  
   $K = $ unweighted Kappa  
   $K_w = $ quadratic weighted Kappa  
   $r = $ Pearson $r$  
   1 = evaluation 1  
   2 = evaluation 2

The first 21 categories in regular typeface received 1-3 ratings. The final eight categories in bold received letter grades of A-D.

The means are across all six rater pairs.

Quadratic weighted Kappas in blue are at or below the accepted level of .60.
Table 5

*Inter-rater Means by Behavior Categories, and Current Rater and Non-current Rater Pairs*

<table>
<thead>
<tr>
<th>Behavior Category</th>
<th>Agr. 1 &amp; 2</th>
<th>Agr. 3 &amp; 4</th>
<th>K 1 &amp; 2</th>
<th>K 3 &amp; 4</th>
<th>K_w 1 &amp; 2</th>
<th>K_w 3 &amp; 4</th>
<th>r 1 &amp; 2</th>
<th>r 3 &amp; 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiar greeting</td>
<td>0.83</td>
<td>0.83</td>
<td>0.71</td>
<td>0.72</td>
<td>0.86</td>
<td>0.88</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>Familiar kennel</td>
<td>0.96</td>
<td>0.75</td>
<td>0.93</td>
<td>0.54</td>
<td>0.95</td>
<td>0.67</td>
<td>0.95</td>
<td>0.74</td>
</tr>
<tr>
<td>Alone outpen</td>
<td>0.71</td>
<td>0.67</td>
<td>0.26</td>
<td>0.00</td>
<td><strong>0.50</strong></td>
<td><strong>0.00</strong></td>
<td>0.59</td>
<td>1.00</td>
</tr>
<tr>
<td>Alone unfamiliar</td>
<td>0.71</td>
<td>0.79</td>
<td>0.32</td>
<td>0.31</td>
<td><strong>0.56</strong></td>
<td><strong>0.26</strong></td>
<td>0.57</td>
<td>0.32</td>
</tr>
<tr>
<td>Doorway</td>
<td>0.71</td>
<td>0.67</td>
<td>0.54</td>
<td>0.31</td>
<td>0.68</td>
<td><strong>0.42</strong></td>
<td>0.73</td>
<td>0.54</td>
</tr>
<tr>
<td>Enter kennel</td>
<td>0.88</td>
<td>0.63</td>
<td>0.78</td>
<td>0.43</td>
<td>0.87</td>
<td>0.74</td>
<td>0.89</td>
<td>0.76</td>
</tr>
<tr>
<td>Inside kennel</td>
<td>0.83</td>
<td>0.71</td>
<td>0.73</td>
<td>0.56</td>
<td>0.85</td>
<td>0.80</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>Leash indoors</td>
<td>0.67</td>
<td>0.50</td>
<td>0.44</td>
<td>0.17</td>
<td>0.74</td>
<td><strong>0.45</strong></td>
<td>0.78</td>
<td>0.47</td>
</tr>
<tr>
<td>Leash outdoors</td>
<td>0.88</td>
<td>0.71</td>
<td>0.75</td>
<td>0.54</td>
<td>0.91</td>
<td>0.76</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>Leashing</td>
<td>0.96</td>
<td>0.67</td>
<td>0.92</td>
<td>0.47</td>
<td>0.97</td>
<td>0.78</td>
<td>0.97</td>
<td>0.82</td>
</tr>
<tr>
<td>Novel object</td>
<td>0.71</td>
<td>0.67</td>
<td>0.42</td>
<td>0.41</td>
<td>0.71</td>
<td>0.61</td>
<td>0.76</td>
<td>0.68</td>
</tr>
<tr>
<td>Novel greeting</td>
<td>0.79</td>
<td>0.75</td>
<td>0.68</td>
<td>0.63</td>
<td>0.84</td>
<td>0.80</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>Novel kennel</td>
<td>0.71</td>
<td>0.67</td>
<td>0.54</td>
<td>0.48</td>
<td>0.72</td>
<td>0.68</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>Out kennel</td>
<td>0.88</td>
<td>0.79</td>
<td>0.81</td>
<td>0.69</td>
<td>0.90</td>
<td>0.86</td>
<td>0.92</td>
<td>0.86</td>
</tr>
<tr>
<td>Outpen handling</td>
<td>0.67</td>
<td>0.83</td>
<td>0.51</td>
<td>0.72</td>
<td>0.78</td>
<td>0.80</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>Person outpen</td>
<td>0.92</td>
<td>0.67</td>
<td>0.87</td>
<td>0.45</td>
<td>0.93</td>
<td><strong>0.60</strong></td>
<td>0.94</td>
<td>0.66</td>
</tr>
<tr>
<td>Person unfamiliar room</td>
<td>0.88</td>
<td>0.67</td>
<td>0.77</td>
<td>0.46</td>
<td>0.88</td>
<td>0.68</td>
<td>0.90</td>
<td>0.71</td>
</tr>
<tr>
<td>Stairs</td>
<td>0.92</td>
<td>0.88</td>
<td>0.85</td>
<td>0.79</td>
<td>0.93</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Toy indoor</td>
<td>0.96</td>
<td>0.83</td>
<td>0.93</td>
<td>0.64</td>
<td>0.96</td>
<td>0.64</td>
<td>0.97</td>
<td>0.68</td>
</tr>
<tr>
<td>Toy outdoor</td>
<td>0.75</td>
<td>0.71</td>
<td>0.59</td>
<td>0.37</td>
<td>0.74</td>
<td><strong>0.50</strong></td>
<td>0.77</td>
<td>0.60</td>
</tr>
<tr>
<td>Unfamiliar handling</td>
<td>0.88</td>
<td>0.75</td>
<td>0.75</td>
<td>0.51</td>
<td>0.77</td>
<td>0.67</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td>Socialization familiar</td>
<td>0.79</td>
<td>0.83</td>
<td>0.68</td>
<td>0.75</td>
<td>0.89</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Socialization unfamiliar</td>
<td>0.67</td>
<td>0.71</td>
<td>0.54</td>
<td>0.60</td>
<td>0.85</td>
<td>0.86</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Leashwalk unfamiliar</td>
<td>0.83</td>
<td>0.42</td>
<td>0.77</td>
<td>0.21</td>
<td>0.93</td>
<td><strong>0.54</strong></td>
<td>0.93</td>
<td>0.59</td>
</tr>
<tr>
<td>Leashing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## INTRA-RATER AND INTER-RATER RELIABILITY

<table>
<thead>
<tr>
<th>Category</th>
<th>Unfamiliar</th>
<th>Kennel Presence Familiar</th>
<th>Kennel Presence Unfamiliar</th>
<th>Handling Unfamiliar</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.75</td>
<td>0.67</td>
<td>0.62</td>
<td>0.51</td>
<td>0.93</td>
</tr>
<tr>
<td>Kennel presence</td>
<td>0.79</td>
<td>0.79</td>
<td>0.67</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>familiar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kennel presence</td>
<td>0.71</td>
<td>0.46</td>
<td>0.57</td>
<td>0.23</td>
<td>0.77</td>
</tr>
<tr>
<td>unfamiliar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handling</td>
<td>0.79</td>
<td>0.79</td>
<td>0.70</td>
<td>0.68</td>
<td>0.92</td>
</tr>
<tr>
<td>unfamiliar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.79</td>
<td>0.71</td>
<td>0.72</td>
<td>0.61</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Note: Agr. = agreement measures
- \( K \) = unweighted Kappa
- \( K_w \) = quadratic weighted Kappa
- \( r \) = Pearson \( r \)

The first 21 categories in regular typeface received 1-3 ratings. The final eight categories in bold received letter grades of A-D.

Raters 1 & 2 are current evaluators. Raters 3 & 4 are non-current evaluators.

Quadratic weighted Kappas in blue are at or below the accepted level of .60.
**Figure 1A.** Rater 1, evaluation 1 & 2, intra-rater ratings for one dog.

**Figure 1B.** Rater 1, evaluation 1 & 2, intra-rater ratings for one dog.
Figure 2A. Ratings for a dog with few concerns vs. a dog with severe concerns.

Figure 2B. Ratings for a dog with few concerns vs. a dog with severe concerns.
**Figure 3A.** Inter-rater quadratic weighted Kappa means by single context behavior categories and current and non-current rater pairs.
INTRA-RATER AND INTER-RATER RELIABILITY

**Figure 3B.** Inter-rater quadratic weighted Kappa means by multiple context behavior categories and current and non-current use rater pairs.
Figure 4A. Inter-rater quadratic weighted Kappa means by single context behavior categories and evaluations.
Figure 4B. Inter-rater quadratic weighted Kappa means by multiple context behavior categories and evaluations.
Appendix A

SOCIABILITY SCALE

1. Solicitation, sustained proximity: Dog stays in same quadrant as person a total of at least 50% of the time AND actively solicits social interaction: looks at person while wagging, leans on person, play bows, paws at person, climbs into person’s lap, nudges person’s hand, playfully mouths on person, bounces and playfully barks at person and/or licks person’s face.

2. Solicitation, short proximity: Dog stays in same quadrant as person a total of < 50% of the time AND actively solicits social interaction: looks at person while wagging, leans on person, play bows, paws at person, climbs into person’s lap, nudges person’s hand, playfully mouths on person, bounces and playfully barks at person and/or licks person’s face.

3. Solicitation, no proximity: Dog does not enter same quadrant as person BUT looks at person while wagging, playfully barking, play bowing and/or wiggling body. These behaviors may be brief or sustained, but are clearly directed at person.

4. No solicitation, sustained proximity: Dog stays in same quadrant as person a total of at least 50% of the time but does not actively solicit social interaction. May show some investigatory behavior towards person (sniff person’s body or clothing), sit or lie down near person and/or show food-seeking behavior (sniffing or pawing at bait bag/pockets, etc.). Does not wag or show other signs of solicitious, social behavior.

5. No solicitation, short proximity: Dog stays in same quadrant as person a total of < 50% of the time but does not actively solicit social interaction. May show some investigatory behavior towards person (sniff person’s body or clothing), sit or lie down near person and/or show food-seeking behavior (sniffing or pawing at bait bag/pockets, etc.). Does not wag or show other signs of solicitious, social behavior.

6. No solicitation, no proximity: Dog does not enter same quadrant as person and does not actively solicit social interaction. May show some investigatory behavior towards person from a distance (air-scenting and/or looking at person), but does not wag or show other signs of solicitious, social behavior.

- “Quadrants” actually take the form of concentric circles marked on the floors of the kennels, outdoor pen and unfamiliar room with tape and represent approximate thirds of the available space.

Note: This scale is used to summarize the dog’s social behavior in each context in the evaluation. (ASPCA Behavioral Rehabilitation Center, copyright 2014, used with permission).
BOLDNESS-FEARFULNESS (PEOPLE, OBJECTS & DOGS)

1. Excited, playful or actively interested

**Body**: Tail wags at or above neutral position. Ears are forward or relaxed. Body is not lowered/crouched and appears relaxed, not tense.

**Behavior**: Dog does not move hesitantly, startle or display signs of fear. Dog is excited about or interested in the stimulus. Dog is playful, actively seeks proximity to or investigates the person/object/other dog.

2. Neutral – Not excited or interested, not fearful

**Body**: Tail wags a bit or is held at or above neutral position. Ears are forward or relaxed. Body is not lowered/crouched and appears relaxed, not tense.

**Behavior**: Dog does not move hesitantly, startle or display signs of excitement or fear. Dog is aware of person/object/other dog but shows little interest in it. Does not show signs of fear and does not actively avoid the stimulus. Dog may sit or lie down in a comfortable position with muscles relaxed, walk or trot around the room, stand and look around, or air scent. May look at or approach the exit once or twice but does not focus on it for the majority of the subtest.

3. Mildly fearful/anxious

**Body**: Body is somewhat or intermittently tense. Tail is low, ears may be back. Posture may be lowered. May sit or lie down with tensed muscles. Dog may assume submissive posture (ears back, low fast tail wag, may lick lips, may lift paw or paw person, may submissively urinate).

**Behavior**: Dog seems aware of person/object/dog and may show some interest but does so from a distance (may be unwilling to approach but may watch or air-scent towards stimuli) or may show approach/avoidance behavior. Shows vigilance, mild hesitation, cautious slow movement and/or brief moments of immobility/inactivity. May startle but recovers quickly. May yawn, lip lick, fidget, pace or scratch self.

4. Moderately fearful/anxious

**Body**: Dog’s body is tense/rigid. Tail is low or tucked, ears are back, and eyes are wide. May crouch, sit in a hunched position, lie down with tense muscles and/or tremble.

**Behavior**: Dog seems aware of but does not approach or actively avoids person/object/dog. Dog may remain immobile or may move with significant/prolonged hesitation.

Alternatively, dog may pace, look for an escape route, or move in a very vigilant, nervous (but not panicked) manner.

May be focused on the exit. May startle, showing poor recovery afterwards, becoming more tense or agitated.
5. Extremely fearful/anxious

**Body**: Dog's body is very tense/rigid. Tail is very low or tucked, ears are back, and eyes are wide. May crouch, flatten body against the ground, tremble or pant heavily.

**Behavior**: Dog may be completely immobile for the majority of the time. May try to hide underneath or behind objects or cram body against wall or into a corner. May seem "stuck" in a position (sitting, standing in place or lying down). May stare into space or appear to "fall asleep," especially when forced into close proximity with the person/object/other dog. May be catatonic (seeming completely unaware or unresponsive, standing, sitting or lying in an expressionless, motionless state).

Alternatively, dog may show active avoidance, pacing, fleeing, or persistently seeking an escape route. May show extreme agitation or vigilance.

May startle, showing no recovery afterwards, becoming more tense or agitated. May scream. May crawl along the ground. May lose control of bladder or bowels, or blow its anal glands. May vigorously scratch at or jump up on the exit door, attempting escape. Person may not be able to complete the activity with the dog.

Note: This scale is used to summarize the dog’s fear or anxiety levels in each context of the evaluation. (ASPCA Behavioral Rehabilitation Center, copyright 2014, used with permission).

**AGGRESSION SCALE**

1. None observed
2. Hard stare, whale eye, rigid, freezes
3. Head whips, lifts lip, growls or alarm barks
4. Air snaps or muzzle punches
5. Tries to bite or bites
0. Subtest skipped/not completed

Note: This scale is used to summarize the dog’s level of aggression in each context of the evaluation. (ASPCA Behavioral Rehabilitation Center, copyright 2014, used with permission).
NOT CONCERNING, RED FLAG & ALERT

Not concerning
Behavior is what would be expected in an average, well-adjusted dog.
- No social behavior is acceptable if it was not being solicited by a person
- Mild fear is acceptable in situations that are unusual or startling (unfamiliar room, unfamiliar person, left alone, novel object, doll)

Red flag
Mild issues that an adoptions counselor would **like** to tell potential adopters about to make them aware, such as:
- An aloof dog
- Mild guarding
- No social behavior when a person tries to solicit it
- Mild fear in situation that would NOT be expected to elicit fear in an average, well-adjusted dog (such as a familiar environment, a familiar person, toys/play, food)

Alert
Moderate to severe issues that an adoptions counselor would **need** to tell potential adopters about because they need special management, treatment or care, such as:
- Moderate or extreme fear
- Aggression

*Note:* Individual context rating scale. This scale is used to subjectively summarize the dog’s behavior in each context encountered during the behavior evaluation. The scale is implemented numerically, with 1 = not concerning, 2 = red flag and 3 = alert. (ASPCA Behavioral Rehabilitation Center, copyright 2014, used with permission).
### Dog Behavior Evaluation Key

**A Dogs** exhibit no significantly concerning behavior.

These dogs show no aggressive behavior during the evaluation, and they show no fear or anxiety for the majority of the evaluation. Some appear nervous during the first few minutes of the evaluation but quickly relax and show no further signs of significant anxiety. Some exhibit mildly fearful responses to a situation or stimulus during the evaluation, but such responses are temporary—the dog recovers in just a few seconds. In our opinion, A Dogs are ready for the adoption floor and require no special treatment or adopter counseling.

**B Dogs** exhibit mild fear, anxiety or aggression.

Some of these dogs show signs of mild fear or anxiety throughout the evaluation. Others show mild fear in response to a certain situation or stimulus and do not recover immediately. Some show mild aggressive behavior, such as growling or freezing in response to a person touching them while they eat or yelping and mouthing or orienting to a person’s hand during uncomfortable handling. However, B Dogs do not bite or attempt to bite during the evaluation.

Some B Dogs may not show any behavior problems once they adjust to a new home. Others would benefit from behavior modification sessions before or after placement. Foster care may be beneficial for some fearful B Dogs. Adopter counseling before placement and/or placement restrictions (no young children, no other dogs, experienced owner, etc.) may be appropriate.

**C Dogs** exhibit moderate fear, anxiety or aggression.

Fearful C Dogs show signs of moderate fear or anxiety for the majority of or during the entire evaluation. If startled, they do not recover—they remain fearful and may attempt to hide or become immobile. Most do show at least some sign of social behavior towards people during the evaluation, however some are undersocialized to people and only show social behavior toward them when in the presence of other dogs.

Aggressive C Dogs may (1) show aggression during the majority of the evaluation, in response to multiple situations or stimuli, (2) bite or attempt to bite at some point during the evaluation or both.

C Dogs may need long-term behavior modification and/or management. Foster care may be necessary for some fearful C Dogs. Adopter counseling before placement is strongly recommended. Placement restrictions (no young children, no other dogs, experienced owner, etc.) are often appropriate.

**D Dogs** exhibit severe fear or aggression.

They fall into two categories: (1) dogs that are too aggressive toward people or other animals to place in an adoptive home or (2) dogs that are suffering from an extremely poor quality of life due to extreme fear or anxiety. If D Dogs fall into the second category, they show no social behavior toward people during the evaluation. They may exhibit feral behavior. In our opinion, prognosis is poor in either case, and euthanasia is the most appropriate outcome.

---

*Note: Multiple context grading scale. This scale is used to subjectively summarize the dog’s behavior in multiple contexts encountered during the evaluation, including in the form of an overall grade. (ASPCA Behavioral Rehabilitation Center, copyright 2014, used with permission).*
Appendix C

How the pieces fit together