Applied Animal Behaviour Science 141 (2012) 139–148

Contents lists available at SciVerse ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim

Investigating behavior assessment instruments to predict aggression in dogs

Sara L. Bennett[a], Annette Litster[a,*], Hsin-Yi Weng[b], Sheryl L. Walker[a], Andrew U. Luescher[c]

[a] *Department of Veterinary Clinical Sciences, College of Veterinary Medicine, Purdue University, Lynn Hall, 625 Harrison Street, West Lafayette, IN 47907, USA*
[b] *Department of Comparative Pathobiology, College of Veterinary Medicine, Purdue University, VPTH, 785 Harrison Street, West Lafayette, IN 47907, USA*
[c] *Fondation Barry, 1928 Martigny, Switzerland*

ARTICLE INFO

Article history:
Accepted 16 August 2012
Available online 5 September 2012

Keywords:
Behavior assessment instrument
Temperament test
Shelter
Dogs
Aggression

ABSTRACT

This masked controlled study evaluated a group of dogs to determine if the results of two behavior assessments detected aggression in dogs that had a history of aggression according to a validated questionnaire for measuring behavior and temperament traits in dogs. Groups of dogs with or without a history of aggression were identified from owner-completed questionnaires for 67 dogs. Any dogs that had a maximum score of no greater than 1 for any question comprising aggression factors were placed in the low/no aggression group and any dogs that had a maximum score of 2 or higher on any question comprising the aggression factors were placed in the moderate to severe aggression group. This second group was further divided to separate moderate aggression from severe aggression. Two behavior assessments, Meet Your Match (MYM)^{TM} Safety Assessment for Evaluating Rehoming^{TM} (SAFER^{TM}) (SAFER) and a modified version of Assess-A-Pet (mAAP), were administered to each dog in random order by assistants masked to the dogs' behavioral histories. The scores for each assessment were divided into binary categorizations (no aggression or aggression). For SAFER, the aggression category was further divided, separating dogs that showed fear, arousal or inhibited aggression from those that showed moderate aggression, and from those that showed severe aggression. The previously established categories for the mAAP of 'no issue', 'unsocial', 'borderline' and 'fail' were also used. Subtest scores for each assessment were also summed. With binary categorization, SAFER showed both lower sensitivity and specificity at 0.60 (95% confidence limits (CL) = 0.44, 0.74) and 0.50 (95% CL = 0.28, 0.72) respectively, than mAAP at 0.73 (95% CL = 0.58, 0.85) and 0.59 (95% CL = 0.36, 0.79) respectively. The odds ratio showed that an aggressive dog was 4.1-fold more likely to be classified in an aggression group by the mAAP test and 1.5 times more likely by SAFER. When the assessments were split into multiple categories, SAFER results were no longer significant, but mAAP maintained a statistically significant but weak correlation of 0.34 ($P = 0.005$) with historical aggression categories. SAFER testing was unable to identify dogs with moderate aggression that could potentially be addressed with behavior modification. By independently selecting acceptable levels of false positive or false negative results for the assessment, summed score results could be used in shelters as an aid to selecting dogs for adoption. Behavioral assessment results should be used in conjunction with other information such as intake history and staff observations to make an informed outcome decision for an individual dog.

1. Introduction

Behavior problems, including aggression, are the most common reasons that dogs are relinquished to shelters

* Corresponding author. Tel.: +1 765 418 3186; fax: +1 765 496 1108.
E-mail address: catvet@purdue.edu (A. Litster).

0168-1591/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.applanim.2012.08.005

(Patronek et al., 1996; Salman et al., 1998). In one study, almost half of dogs were relinquished to shelters due to behavior problems as indicated by the relinquisher and 12% of those dogs had bitten a person (Salman et al., 1998). Additionally, shelters were the second most common source of animals that were subsequently relinquished, with 22.8% of 3676 relinquished dogs having been obtained from a shelter (Salman et al., 1998). Patronek et al. (1996) also noted that dogs acquired from a shelter were at an increased risk of relinquishment. To maximize the success of rehoming relinquished dogs, shelters need to identify relinquished animals that are at particular risk of showing behavior problems, including aggression, if adopted into a new home.

In its simplest form, a temperament test evaluates a behavioral response to a specific stimulus at one time in one environment by an individual dog. Test results are then used to predict how that same animal may respond in a similar situation based on its response in the testing situation (Taylor and Mills, 2006). It is frequently assumed that the behavior will be consistent over time, and that the dog's temperament or personality is being measured. Many experts prefer to call these tests 'behavior assessments' due to the lack of published evidence to confirm that they predict stable behavior over time, and therefore whether test results are truly associated with temperament or personality traits.

The goals of these tests in a shelter setting are firstly to predict an aggressive response, which would then allow the shelter to prevent placing potentially dangerous dogs in new homes, thereby mitigating a major public safety concern, and secondly to identify dogs with potentially treatable or manageable behavior problems. Test results can then be used to help make informed adoption decisions or to make recommendations for rehabilitation where possible (Diederich and Giffroy, 2006; Jones and Gosling, 2005). An overarching goal is to successfully match potential owners to suitable dogs so that long-term homes are achieved.

In shelters, tests typically applied are in a test battery format. This is defined as a 'series of standardized experimental situations where stimuli serve to elicit behavior that is then statistically compared to other individuals placed in the same situation in order to classify the subject tested' (Diederich and Giffroy, 2006). Most of these tests fall into two different categories – they either simulate the average home environment or interactions typical of those with an average owner, or they attempt to elicit aggression through intensely stressful situations (Segurson, 2009).

While they can be useful tools, there are some inherent problems with the practical application of behavior assessments in shelters. Previous studies have raised concerns that some types of behavior, specifically intra-species aggression, predatory behavior, and territorial aggression, may not be accurately identified in a shelter setting (Christensen et al., 2007). Some behavior assessments were created to identify different types of aggression in populations of dogs with a high prevalence of aggression (Netto and Planta, 1997) and may be considered extremely provocative to the average dog, resulting in an unacceptably high percentage of false positive results in a shelter setting. Additionally, several well known, commonly used, heavily researched tests are complex and time consuming (Netto and Planta, 1997; van der Borg et al., 1991, 2010) and therefore impractical for use in shelters with limited resources. Because of these factors, behavior testing in shelters often consists of untested and unvalidated combinations of subtests originating from various testing instruments. Of particular concern is the lack of standardization between shelters and between staff members within shelters (Mornement et al., 2010). Another limiting factor is that in the test validation process, researchers are unable to include those dogs that show aggression on the assessments because most of them are not placed for adoption, thereby losing the opportunity for follow up to evaluate the test for false positive results (Bollen and Horowitz, 2008). False identification of aggression results in dogs being needlessly condemned, while failure to identify aggression poses a public safety risk (Segurson, 2009). Although it is ideal to evaluate dogs with historical aggression by investigating specific incidents and possible environmental and situational triggers on an individual basis, this is often not possible in a shelter setting due to the incomplete nature of behavioral histories and the unfamiliar and possibly stressful shelter environment. Most tests used in shelters attempt to replicate common interactions or situations that dogs would encounter in an average home, and these assessments include evaluations for concerns with handling, rough play, food or other resource guarding, exposure to strangers, children, other dogs or animals. The internal validity of these situations is debatable, and the results of some subtests might be affected by the dog reacting with fear or aggression to the stimulus of an unfamiliar person, thereby invalidating the intended purpose of the subtest. Additionally, it is possible that the intended trigger might not be specific enough to elicit the behavior of concern.

The aim of this masked controlled study was to evaluate a group of dogs for which a behavioral history was available, using a validated questionnaire (the Canine Behavioral Assessment & Research Questionnaire – C-BARQ), to determine the ability of two behavior assessment instruments commonly used in American shelters to predict aggression – the American Society for the Prevention for Cruelty to Animals (ASPCA) Meet Your Match (MYM^{TM}) Safety Assessment for Evaluating Rehoming^{TM} (SAFER^{TM}) (SAFER), Weiss, 2007 and a modified version of Assess-A-Pet (mAAP) (Bollen and Horowitz, 2008), abbreviated to SAFER and mAAP for the remainder of this paper.

2. Materials and methods

2.1. Ethical approval

This project was approved by the Purdue University Animal Care and Use Committee (Approval # 10-062).

2.2. Subjects

Seventy-three dogs were recruited from the Purdue University Veterinary Teaching Hospital (PUVTH) Behavior Service or Community Practice using a convenience sampling method during a period from August 26, 2010 through June 28, 2011. Inducements were not provided to clients who enrolled their dogs in the study. Dogs were eligible for inclusion if they were between 1 and 8 years old, surgically altered, were currently vaccinated for rabies, and had lived with the current owner for at least 3 months. Since most dogs from US shelters are spayed or neutered prior to adoption, surgical alteration was added as an inclusion criterion to increase the external validity of the study.

2.3. Behavior history

2.3.1. Behavior history questionnaire

Each owner completed a research consent form and a previously validated owner questionnaire for measuring behavior and temperament traits in dogs (C-BARQ: Hsu and Serpell, 2003) to obtain a behavioral history for each dog at the time of the behavioral assessment.

The C-BARQ is an owner-completed questionnaire that has been previously validated for measuring behavior and temperament traits in pet dogs. The following factors have been formally validated: stranger directed aggression, owner directed aggression, dog directed aggression, dog directed fear, stranger directed fear, nonsocial fear, separation related problems, and attachment/attention seeking (Hsu and Serpell, 2003). The C-BARQ has been suggested to be used as a screening tool to identify behavior problems or as a tool to measure the efficacy of behavioral treatment plans. The data from this questionnaire were used as the reference standard to identify dogs with or without a history of aggression so that enrolled dogs could be categorized. All of the previously-validated aggression factors from the C-BARQ, including stranger-directed aggression, owner-directed aggression, and dog-directed aggression were used to create the test groups. Details of the triggers and contexts included in the questions for each factor have been described previously (Hsu and Serpell, 2003; Segurson et al., 2005). Each C-BARQ factor was comprised of 4–10 questions, each scored using a five-point behavior scale, with ordinal scores of 0–4 representing escalating severity of aggression. A score of 0 indicated no aggression; a score of 1, 2 or 3 indicated mild to moderate aggression, such as barking, growling, and/or showing teeth, on an increasing scale. A score of 4 indicated serious aggression identified as snapping, biting, or attempting to bite.

2.3.2. Groups based on behavior history

With each factor, dogs were grouped based on the highest score for that factor. These groups were subdivided as follows – a maximum score of 0 or 1 on any question comprising the factor was noted with a ‘0’; a maximum score of 2 or 3 on any question comprising the factor was noted with a ‘1’; and a maximum score of 4 on any question comprising the factor was denoted by a ‘2’. Each dog was scored for each of the three factors used. For the purposes of further analysis, each dog was then categorized in the following two ways.

Any dogs that had all three factor groups falling in the ‘0’ category were designated to group ‘0’ indicating that the dog had a history of no or possibly mild aggression. Any dogs that had at least one factor score of ‘1’ or ‘2’ were placed in group ‘1’, indicating the dog had a history of moderate or severe aggression.

The aggression group was further categorized into moderate aggression and severe aggression. Any dog that had a maximum factor score of ‘1’ was placed in group ‘1’, indicating moderate aggression, and any dog that had a maximum factor score of ‘2’ was placed in group ‘2’, indicating severe aggression.

2.4. Behavior assessments

2.4.1. Behavior assessment procedures

Two temperament tests, SAFER (Weiss, 2007) and mAAP previously investigated by Bollen and Horowitz (2008), were administered to each dog by two members of a team of four volunteers masked to the dogs’ behavioral history. These tests were chosen for the investigation because they are commonly used by shelters in the United States, either as written, or as modified versions. The tests were administered to each dog in random order using a random number table (Steel and Torrie, 1960). Data were recorded using video recording and written notes. SW, who was SAFER certified through the ASPCA MYM^{TM} SAFER^{TM} Certification program, handled the dogs as the evaluator and performed 70 (93.3%) of the assessments. She later reviewed video records of the seven remaining assessments (6.7%) to ensure that the tests were performed exactly as written.

2.4.2. SAFER assessment

This assessment was made up of seven different subtests as summarized below, performed in the following order as written. The equipment used was as specified by the training manual, including a buckle collar and 1.8 m leash, a fake plastic hand on a dowel rod, a rope toy, plastic squeak toy, plush squeak toy, plain un-basted rawhide, and a mixture of canned and dry dog food.

1. Look – Evaluator gently held dog’s head in her hands and looked into the dog’s eyes, with non-threatening eye contact.
2. Sensitivity – Evaluator gently grasped handfuls of fur and skin in a firm kneading motion over the dog’s side from shoulder to hip and back two times.
3. Tag – Evaluator tried to engage the dog using an excited voice and play movement, then touched the dog lightly with a finger to try to initiate play. This was repeated twice.
4. Squeeze – Evaluator said “squeeze”, then ran a hand down the dog’s leg, then gently squeezed between the toes using the pads of her fingers, and repeated.
5. Food behavior – The dog was given a mix of canned and dry food and allowed to start eating. The evaluator then asked for the food and placed a fake hand into the food bowl, pulling the bowl away from the dog, and allowed the dog to begin eating again. The evaluator then pushed the dog’s head away from the food bowl gently or stroked the dog’s head and neck using a fake hand.
6. Toy behavior (with optional rawhide behavior) – Evaluator showed the dog a toy and then tossed it to the dog and gave it time to take possession. The evaluator then reached in to take the toy using a fake hand. If the dog

showed no interest in the toy, a different toy was offered in the same manner. The same procedure was repeated with a plain un-basted rawhide.

7. Dog-to-dog behavior – The test dog entered the testing room on leash with the evaluator where the observer waited with a stable non-reactive dog on leash. The dogs did not touch or greet each other, and only the initial approach was evaluated.

2.4.3. *mAAP assessment*

This assessment was made up of nine subtests previously described (Bollen and Horowitz, 2008) and summarized below. They were performed in the following order as written. Equipment used was as specified in mAAP as written, including a heavy cotton English slip lead, towel, a fake plastic hand on a dowel rod, a rope toy, plastic squeak toy, plush squeak toy, beef basted pig's ear, a mixture of canned and dry dog food, and jacket and hat for the 'stranger' to wear during the stranger subtest.

1. Cage presentation – The evaluator stood quietly for 5 s in front of cage, then gave 5 s of non-threatening eye contact, then knelt down and talked to dog in a friendly manner for 5 s.
2. Sociability – The evaluator stood and ignored dog for 60 s while holding the leash, then stroked the dog's back three times, then sat down and ignored dog for 5 s, then talked to the dog in a friendly manner for 20 s.
3. Teeth examination – The evaluator lifted the dog's upper lips and held them for 5 s, and attempted to repeat this five times.
4. Handling – The evaluator stroked the dog's side, picked up a hind foot, tugged on the tail, touched both ears, wiped the dog's body with a towel, tugged on the collar, pressed on the dog's shoulders, and hugged the dog.
5. Arousal – The evaluator initiated play using toys with the dog for 30 s then stopped.
6. Food bowl – While the observer held the dog's leash, the evaluator gave the dog a bowl of canned and dry food and used a fake hand to pet the dog's back, then reach toward the bowl and push the dog's face away. This was repeated three times.
7. Possessions – While the observer held the dog' leash, the evaluator gave the dog a basted pig's ear and then said 'drop it' and attempted to take it away with the fake hand.
8. Stranger – A person dressed in a hat and coat knocked on the door, entered, and then walked toward the dog and made eye contact for 3 s, then stepped forward and reached toward the dog. The stranger then knelt down and talked to the dog in a friendly manner.
9. Dog introduction – The test dog was presented with an unfamiliar neutral dog on a leash.

2.5. *Testing environment*

All portions of the behavior assessments were performed in a windowless room measuring 4.4 m by 5.8 m. The flooring was concrete and the exit from the room was blocked by an exercise pen. A large dog crate or the exercise pen was used for the cage presentation subtest of mAAP, depending on the size of the dog. An English slip lead was used to maintain control of the dog during mAAP, and a buckle collar and nylon leash 1.8 m in length were used during SAFER assessments, as directed by each respective test guidelines. All other supplies were used as directed according to the respective authors' specifications (Bollen and Horowitz, 2008; Weiss, 2007). One team member handled the dog to perform the assessments as written (evaluator) and the other team member set up the video recorders, recorded scores, and assisted during the assessments as designated (observer). The video recorders used were a Kodak Zi8 pocket video camera (Sylmar, CA, USA) recording in real time at 1080p (1920 × 1080) at 30 frames/s and a Panasonic DIGA DVD Palmcorder® MultiCam^{TM} Camcorder VDR-D100 (Seacaucus NJ, USA) recording in real time at 29.95 frames/s. If unanticipated noises or distractions occurred during the assessment, the assessment was paused and resumed once the distraction was no longer present.

2.6. *Categorization of data*

2.6.1. *Categorization of SAFER data*

In the SAFER assessment scoring sheet, each subtest was scored using a series of behavioral descriptions related to a scale of 1–5, with higher numbers indicating escalating signs of aggression. Certain scores on each subtest could then be categorized using a legend (Weiss, 2007). Using this legend (Weiss, 2007), a designation of 'P' would correspond to a numeric score of 3 (or 4 on the dog-to-dog subtest), and indicated some concerning behaviors, such as signs of fear, high arousal or inhibited aggression, such as stiffening or whipping the head around. From this it was suggested that the dog may benefit from potential behavior modification or management. A designation of 'R' denoted that the dog displayed some mild to moderate aggression, corresponding to a score of 4 on all but the dog-to-dog subtest, or a score of 5 on the second squeeze subtest and all later subtests. This indicated that behavior modification or management was strongly recommended, while a designation of 'S' signaled a need to stop the assessment for safety reasons. This designation correlated to a score of 5 on the first four subtests (Weiss, 2007). For the purposes of further analysis, SAFER assessment results were categorized in two ways as follows:

(1) Any dog that did not receive any 'P's, 'R', or 'S's through all subtests was placed in a category of '0', indicating no aggression or concerning behavior was noted. Any dog that received at least one score that corresponded to a 'P', 'R', or 'S' was placed into a category of '1'.

(2) Any dog that did not receive any 'P's, 'R', or 'S's through all subtests was placed in a category of '0', indicating no aggression was noted. Any dog that received at least one score that corresponded to a 'P', but no dog that received an 'R' or 'S' was placed into a category of '1'. Any dog that received at least one score that corresponded to an 'R' was placed into a category of '2'. Any

dog that received at least one score that corresponded to an 'S' was placed into a category of '3'.

2.6.2. Categorization of mAAP data

Each subtest was scored by choosing from a series of behavioral descriptions. Descriptions that were bolded, indicating aggression or other undesirable behavior, such as lack of interest in evaluator during the sociability subtest, were scored as a fail for that subtest. As Bollen and Horowitz (2008) noted, the scores from their assessments were then used to separate the dogs into four categories. Dogs in the 'no issue' group showed no aggression or other concerning behaviors during the assessment. Dogs in the 'unsocial' group did not show aggression during any subtest, but did not show interest in social interaction during the sociability subtests. Dogs in the 'borderline' group showed mild aggression such as becoming stiff or some growling in no more than two subtests. Dogs in the 'fail' group showed mild aggression in more than two subtests or showed severe aggression such as intense growling, lunging or attempting to bite in any one subtest. For statistical analysis, these groups were used to categorize the dogs in two ways as follows:

(1) Any dogs that fell into the 'no issue' or 'unsocial' group were designated with a '0', indicating no aggression noted during the assessment. Any dogs that fell into the 'borderline' or 'fail' group were designated with a '1', indicating aggression was noted during the assessment.
(2) The dogs in the 'no issue' group were designated with a '0', the 'unsocial' group were designated with a '1', the 'borderline' group were designated with a '2', and the 'fail' group were designated with a '3'.

2.6.3. Subtest summed scores

For each behavior assessment, the score for each subtest was summed for each dog. For SAFER, the numeric ordinal score for each subtest was summed so that a dog could receive a minimum score of 9 (no aggression) and a theoretical maximum score of 45 (severe aggression in all subtests). For mAAP, each bolded behavioral description, indicating 'failure' of the subtest, was designated with a numeric score of '1', while any non-bolded description was designated with a numeric score of '0'. The numeric scores for each subtest were then summed so that there was a minimum score of 0 (no aggression) and a maximum score of 9 (aggression in all subtests). There were several dogs that did not have a score for all subtests. One reason for this was that the dog showed no interest in food. The scores for the subtests with food were therefore eliminated from the calculation of the summed score for each dog. SAFER specified that if the dog showed no interest in the rawhide or toy, however, it received a score of '1' for that subtest, thus identifying the interpretation of 'no interest', depending on whether the item of interest was food or a 'possession' such as a toy or rawhide. There was no such identification made in the scoring choices for mAAP, however. The decision to leave the subtest score blank was made with the food and rawhide (possession) subtests for mAAP assessment because the test instructions did not specify how to score a dog that showed no interest in the item. Another reason for a blank subtest score was that if the dog showed serious aggression, the handling portion of the assessment was discontinued for safety reasons. Although it was understood that discontinuing interaction that was resulting in aggression could potentially teach the dog that aggression could be used to stop a stressful interaction, the safety of the handlers and the welfare of the dog was considered of higher importance at that point. It was concluded that if scored, those subtests would likely have a higher score, and the corresponding subtest was given a score of 5, indicating growling, lunging or attempting to bite, for SAFER and a score of 1, indicating stiffening, whale eye, growling, snapping, or attempting to bite, for mAAP.

2.7. Statistical analysis

We used sensitivity, specificity, false positives, false negatives, and odds ratios to evaluate and compare the performance of SAFER and the mAAP. The test results in binary categorization were compared to dogs with or without a history of aggression, using the behavioral history from the C-BARQ questionnaire completed by the dog's owner, as the reference standard. When considering dogs with a history of aggression, test sensitivity was defined as the probability that the test would accurately identify a dog with a history of aggression. It was calculated by dividing the number of dogs who have a history of aggression and were correctly classified in an aggression category using the assessment tool by the total number of dogs with a history of aggression. False negatives were computed as the number of dogs that had a history of aggression and were incorrectly classified in a no aggression category using the assessment tool divided by the total number of dogs with a history of aggression. We computed specificity by dividing the number of dogs that did not have a history of aggression and were classified in a no aggression category using the assessment tool by the total number of dogs with a history of no aggression. Therefore, test specificity was defined as the probability that the assessment accurately identified dogs with a history of no aggression. False positives were computed as the number of dogs that had a history of no aggression and were incorrectly classified in an aggression category using the assessment tool divided by the total number of dogs with a history of no aggression (Feinstein, 2002). Spearman's rank correlation coefficient was used to measure the direction and strength of the correlations between the three-category aggression history and the four-category test results of the two assessments. For summed scores, we performed receiver operating characteristic (ROC) analyses to compare the performance of the two behavior assessment tests using area under the curve (AUC) measurements and to assist in identifying optimal cutoff values. Sensitivity, specificity, likelihood ratio positive (LR+) and negative (LR−) for each cutoff value were provided. The corresponding 95% confidence limits (CL) were calculated. We defined significance as $P < 0.05$. All analyses were performed using Stata/IC 11.2 for Windows (StataCorp LP, College Station, TX).

Table 1
Demographic information and aggression groups (multiple) based on behavior history from the behavior questionnaire (C-BARQ) completed by owners at time of assessments.[a,b]

Characteristic	Number (%)	No to mild aggression (% total sample) (% of characteristic)	Moderate aggression (% total sample) (% of characteristic)	Severe aggression (% total sample) (% of characteristic)
Sex				
Male	33 (49)	9 (13) (27)	14 (21) (42)	10 (15) (30)
Female	34 (51)	12 (18) (35)	8 (12) (24)	14 (21) (41)
Age				
1–2 years	21 (31)	8 (12) (38)	6 (9) (29)	7 (10) (33)
3–4 years	19 (28)	5 (7) (26)	5 (7) (26)	9 (13) (47)
5–6 years	16 (24)	5 (7) (31)	6 (9) (38)	5 (7) (31)
7–8 years	11 (16)	3 (4) (27)	5 (7) (45)	3 (4) (27)
Shelter history				
Yes	28 (42)	4 (6) (14)	12 (18) (43)	12 (18) (43)
No	39 (58)	17 (25) (44)	10 (15) (26)	12 (18) (31)
SAFER performed 1st				
Yes	34 (51)	11 (16) (32)	14 (21) (41)	9 (13) (27)
No	33 (49)	10 (15) (30)	8 (12) (24)	15 (22) (46)
Total	67 (100)	21 (31) (NA)	22 (33) (NA)	24 (36) (NA)

[a] $n = 67$.
[b] Not all percentages add to 100 due to rounding error.

3. Results

3.1. Test performance and scoring

Of 73 dogs recruited, six were excluded, leaving 67 dogs for analysis. For four of the six dogs, tests were not performed as written; another dog had extensive pharmacologic therapy to control behavior problems and epileptic seizures during the study period; and the remaining dog was excluded because of the incomplete nature of the C-BARQ questionnaire and a subsequent inability to make contact with the owner.

3.2. Descriptive statistics

The age and sex characteristics of the sampled group were evenly distributed when considering the groups divided in a binary fashion or with multiple aggression groups (Table 1). Twenty-one dogs (31%) of the sampled group had been placed in the no or possible mild aggression group (0) based on the C-BARQ questionnaire. Although a higher percentage of dogs were placed in the moderate or severe aggression group (1 or 2), the variables age, sex, and order of assessment performance were evenly distributed between the two aggression groups. Twenty-eight (42%) of the dogs had been obtained from a shelter or rescue group. Twenty-four of those 28 dogs (86%) fell into the moderate or severe aggression group (Table 1).

3.3. Binary assessment category comparison

When the behavior assessments, using binary categorization, were compared to the binary groups based on history (no/mild aggression vs. moderate/severe aggression), SAFER showed both lower sensitivity and specificity than mAAP (Table 2). The results also showed that an aggressive dog (moderate to severe aggression) was 1.5-fold and 4.1-fold more likely to be classified in an aggression group (P, R or S) by SAFER and (borderline or fail) by mAAP tests, respectively, than a dog with no history of aggression. LR+ (and LR−) were 1.2 (0.8) for SAFER and 1.79 (0.45) for mAAP.

Table 2
Sensitivity, specificity, false positives, false negatives, and odds ratio with 95% confidence limits (CL) for aggression based on binary categories from assessment results.[a,b]

	SAFER	mAAP
Sensitivity (95% CL)	0.60 (0.44, 0.74)	0.73 (0.58, 0.85)
Specificity (95% CL)	0.50 (0.28, 0.72)	0.59 (0.36, 0.79)
False positives (95% CL)	0.50 (0.28, 0.72)	0.41 (0.21, 0.64)
False negatives (95% CL)	0.40 (0.26, 0.56)	0.27 (0.15, 0.42)
Odds ratio[c] (95% CL)	1.5 (0.5, 4.2)	4.1 (1.4, 11.7)

[a] Aggression history (moderate to severe vs. no to mild) was based on the behavior questionnaire (C-BARQ).
[b] $n = 67$.
[c] Odds ratio of the dog being classified in an aggression group by the assessment test if the dog had a history of moderate to severe aggression.

3.4. Multiple assessment (four) category comparison

There was a stronger positive association with the binary aggression history when mAAP test was split into four categories than when the SAFER test was split into four categories (Table 3). There was not a significant association between the four-category SAFER results and the binary aggression history ($P = 0.653$) or the multiple group aggression history ($P = 0.078$; Tables 3 and 4). The four-category mAAP test results showed a positive association with the binary aggression history ($P = 0.010$; Table 3) and with the multiple group aggression history ($P = 0.005$; Table 4).

3.5. Summed score comparison

3.5.1. Summed score computation

The protocol for calculating the summed scores as noted in Section 2.6.3 resulted in summed score results from 67 dogs for SAFER and summed score results from 61 dogs for mAAP test.

Table 3
Cross-tabulation between the binary aggression history (moderate to severe vs. no to mild) based on the behavior questionnaire (C-BARQ) and the four-category assessment results.[a]

Assessment	Category[b]	Aggression history		Odds ratio[c]	P-value[d]
		Moderate to severe aggression (%)	No to mild aggression (%)		
SAFER	0	18 (40)	11 (50)	–	0.653
	1	14 (31)	5 (23)	1.7	
	2	4 (9)	2 (9)	1.2	
	3	9 (20)	4 (18)	1.4	
mAAP	0	7 (16)	10 (46)	–	0.010
	1	5 (11)	3 (14)	2.4	
	2	26 (58)	7 (32)	5.3	
	3	7 (16)	2 (9)	5	

[a] $n = 67$.
[b] Higher score indicated higher degree of aggression based on assessment scores.
[c] Odds ratio of the dog being classified as moderate to severe aggression using category 0 as reference group.
[d] Chi-square test for trend.

Table 4
Correlation between the three-category aggression history (moderate to severe vs. no to mild) based on the behavior questionnaire (C-BARQ) and the four-category assessment results.[a]

Assessment	Category[b]	Aggression history			Correlation[c]	P-value
		Severe aggression (%)	Moderate aggression (%)	No to mild aggression (%)		
SAFER	0	6 (25)	12 (55)	11 (52)	0.217	0.078
	1	9 (38)	5 (23)	5 (24)		
	2	2 (8)	3 (14)	1 (5)		
	3	7 (29)	2 (9)	4 (19)		
mAAP	0	3 (13)	4 (18)	10 (48)	0.342	0.005
	1	3 (13)	2 (9)	3 (14)		
	2	13 (54)	13 (59)	7 (33)		
	3	5 (21)	3 (14)	1 (5)		

[a] $n = 67$.
[b] Higher score indicated higher degree of aggression based on assessment scores.
[c] Spearman's rank correlation coefficient.

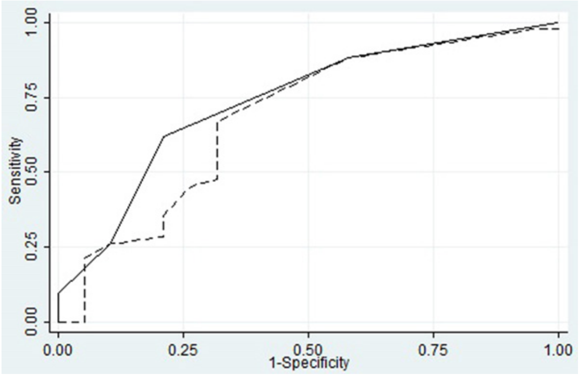

Fig. 1. ROC curve of summed scores for SAFER (dashed line, $n = 67$) and mAAP (solid line, $n = 61$).

3.5.2. *ROC analysis*

The area under the curve for SAFER was 0.68 (95% CL = 0.53, 0.84). The area under the curve for mAAP was 0.74 (95% CL = 0.60, 0.87). The AUC for each of the two assessments were not significantly different from one another ($P = 0.405$; Fig. 1). Sensitivity, specificity, and likelihood ratios for each summed score cutoff point is presented in Table 5 for SAFER ($n = 67$) and Table 6 for mAAP ($n = 61$).

4. Discussion

Both the SAFER and mAAP tests had a marginal sensitivity and specificity, although mAAP more accurately classified aggressive dogs, as evidenced by the odds ratio reported in Table 2 (4.1; 95% CI 1.4, 11.7). Regardless of which assessment tool was used, both false positive and false negative assessment results occurred, even when the assessments were split into more detailed categories. After refined categorization, SAFER results were no longer significantly correlated with the aggression history, but mAAP maintained a statistically significant and positive association with aggression history (Tables 3 and 4). As the assessment score category increased (2 and 3), the percent of dogs with a history of moderate to severe aggression increased, indicating the test could identify dogs in the unsocial category (1) as non-aggressive using the mAAP test. In this study, SAFER testing was unable to identify dogs with moderate aggression that could potentially be addressed with behavior modification (Table 3).

The results from the ROC analysis (Tables 5 and 6) could aid in determining an acceptable cutoff point for the continuous summed scores in an individual shelter, based on the shelter intake population and the relative importance of reducing false positive results (by setting the cutoff point to increase specificity) or false negative results (by setting the cutoff point to increase sensitivity) for the assessment.

Table 5
Sensitivity (SE), specificity (SP), positive and negative likelihood ratio (LR+, LR−) of identifying aggression for different cutoff points of the summed scores from the SAFER assessment.[a,b].

Cutoff	SE (%)	SP (%)	LR+[c]	LR−[d]
≥6	100	5	1.05	0
≥7	98	5	1.02	0.49
≥8	98	9	1.08	0.24
≥9	89	41	1.50	0.27
≥10	69	64	1.89	0.49
≥11	51	64	1.41	0.77
≥12	49	68	1.54	0.75
≥13	40	73	1.47	0.83
≥14	33	73	1.22	0.92
≥15	31	82	1.71	0.84
≥16	27	91	2.93	0.81
≥18	20	91	2.20	0.88
≥26	18	91	1.96	0.90
≥27	11	91	1.22	0.98
≥28	7	95	1.47	0.98
≥29	2	95	0.49	1.02

[a] $n = 67$.

[b] Behavior history based on questionnaire (C-BARQ) was used as a reference standard for aggression. Positive likelihood ratio indicates how much the odds of showing aggression after adoption would increase if a dog showed aggression on the assessment (increasing LR+ value beyond one indicates greater odds of post-adoption aggression). Negative likelihood ratio (LR−) indicates how much the odds of showing aggression after adoption would decrease if the dog did not show aggression on the assessment (decreasing LR− closer to zero indicates reduced odds of post-adoption aggression) (Feinstein, 2002).

[c] LR+ = SE/(1 − SP).

[d] LR− = (1 − SE)/SP.

Table 6
Sensitivity (SE), specificity (SP), likelihood ratio positive (LR+), and likelihood ratio negative (LR−) of identifying aggression for different cutoff points of the summed scores from the mAAP assessment.[a,b].

Cutoff	SE (%)	SP (%)	LR+[c]	LR−[d]
≥0	100	0	1.00	–
≥1	88	42	1.52	0.28
≥2	62	79	2.94	0.48
≥3	26	89	2.49	0.82
≥4	10	100	–	0.90
≥5	2	100	–	0.98

[a] $n = 61$.

[b] Behavior questionnaire (C-BARQ) was used as a reference standard for aggression. Positive likelihood ratio indicates how much the odds of showing aggression after adoption would increase if a dog showed aggression on the assessment (increasing LR+ value beyond one indicates greater odds of post-adoption aggression). Negative likelihood ratio (LR−) indicates how much the odds of showing aggression after adoption would decrease if the dog did not show aggression on the assessment (decreasing LR− closer to zero indicates reduced odds of post-adoption aggression) (Feinstein, 2002).

[c] LR+ = SE/(1 − SP).

[d] LR− = (1 − SE)/SP.

For example, if the avoidance of false negative results (dogs likely to show aggression after adoption that do not show aggression on the assessment) is of primary importance, sensitivity should be increased. Likelihood ratios positive (LR+) and negative (LR−) can also be used to aid in determining optimal cutoff values. LR+ indicates how much the odds of showing aggression after adoption would increase if a dog showed aggression on the assessment (increasing LR+ indicates greater odds of post-adoption aggression) while LR− indicates how much the odds of showing aggression after adoption would decrease if the dog did not show aggression on the assessment (decreasing LR−, indicates reduced odds of post-adoption aggression; Feinstein, 2002).

Summed scores from subtests need to be interpreted with care, recognizing the strengths and limitations of this type of simplification. A major limitation of summed scores is the missing values that can occur if some dogs do not complete all subtests for an assessment. In this study, certain assumptions were applied to address these missing values (Section 2.6.3). For example, we excluded the food subtest in computing the summed score for both SAFER and mAAP assessments, as approximately 85% of tested dogs had at least one missing value for this subtest in both assessments due to the lack of interest in food during the tests. These subtests were not included in the summed score computation because we were unable to retest the dog for food or possessive aggression 24–48 h later and it may be highly inaccurate to label a dog as disinterested in guarding food if other factors precluded the dog from showing interest in food. Therefore, if a dog shows no severe aggression resulting in discontinuation of the assessment, and does not show interest in food (and possibly possessions), a summed score leaving out those subtests can still be calculated. In this way, whether the dog fell above or below that shelter's chosen cutoff point could still be determined based on the scores from other subtests.

This study used privately-owned dogs to investigate behavior assessments that are most commonly used in US shelters. There is a large need for accurate behavioral assessments of shelter dogs since the proportion of the canine pet population acquired through shelters is substantial. American Pet Products Association's annual survey for 2011 reported that 21% of 544 dog owners obtained their dog from shelters or humane societies and 7% obtained their dog from rescue groups (APPA, 2010). Additionally, in a recent US survey of 500 dog owners from 43 Chicago zip codes, 33% of dogs were acquired from shelters (Litster, pers. obs.). Salman et al. (1998) reported that 23% of 3676 dogs relinquished to shelters had originally been obtained from shelters and 82% of the historically aggressive dogs in that study had been obtained from shelters or rescues. Consequently, any study investigating behavior assessments for canine aggression will have inherent relevance to shelter dogs, even if client-owned dogs are used.

The environment, context, and trigger play important roles which can determine whether an individual dog might show aggression in a given situation and it is vital to collect this type of information from relinquishing owners. However, information volunteered by relinquishers might be inaccurate (Segurson et al., 2005) and is unavailable for stray dogs. Additionally, many shelters do not have the resources to address behavior problems in relinquished dogs in order to improve their candidacy for adoption, and decisions regarding the selection of dogs for adoption are often limited by the availability of space and labor. These groups need to be able to identify dogs that have a potential to show aggression in the home to prevent them from being placed for adoption. Other shelters may have access to canine behavior resources and would be willing to try to address some mild to moderate behavior problems to increase adoptability, and need to be able to identify this subset of dogs with manageable or rehabilitatable problems from dogs with severe behavior problems. The binary categorizations used in this study were created to evaluate the ability and strength of the behavior assessments to predict whether dogs would show aggression, in general, once placed in a home. The multiple categories were created to determine if the strength of the test would remain the same or become weaker if it was used to identify varying degrees of overall aggression. If the assessments could be used to predict the severity of aggression, shelters that were equipped to do so would be more able to make informed choices on which dogs they should focus behavior modification resources.

The mAAP test consistently performed better than SAFER through all analyses. However, as mAAP is still prone to error, the results of this behavior assessment test should not be used as the only tool in the decision making process. In order to arrive at a more complete picture of a specific animal, it should be used together with other relevant information concerning the dog. Ideally, screening should be performed on dogs with a higher likelihood of aggression, but how do we identify these dogs? It has been shown in a previous study that staff observations better predict the absence of a problem than the presence of a problem, showing better negative predictive value (van der Borg et al., 1991). Ideally, that information, in conjunction with any available previous history, could be used to determine which dogs should be evaluated with behavior assessments. But, it is often impractical or unsafe to wait until the staff is confident that there are no concerns before assessing the dog. This could endanger staff members as they handle dogs with unidentified aggression problems and also uses valuable time, space and resources for dogs that may not be salvageable by that organization. Intake interviews and relinquishing owner's history can be helpful, but care must be taken to collect this information in a compassionate and private manner, as Segurson et al. (2005) showed that some types of aggression were reported less frequently when the relinquishing owner was told that the information would be used to help make a decision of the dog's future outcome. Judicious use of a variety of information sources should be combined to create a global picture of the dog's behavior and then used to make an informed decision on the dog's outcome.

The goal of this study was to determine if two behavior assessment tools that are frequently used in US shelters were able to identify overall aggression in dogs, but there are important factors that could potentially alter test results in a shelter environment. There is concern for the strength of internal validity of behavior assessments in a test battery format performed on dogs in a shelter setting. Although there is strong opinion as to what exactly each subtest is intended to measure, there is little published evidence to confirm this. The performance of the assessments by a stranger to the dog could easily confound many subtests if the dog has intense fear of strangers and results in aggression due to this. The dog's reaction may be based more on the presence of the stranger rather than the intended trigger comprising the subtest. Additionally, there may be different behavior factors, such as non-social fear, that may be more important to a particular subtest than we anticipate. And, there may be some subtests that have a strong association with a particular behavior factor while other subtests have little to no association with any factors measured. This presents a significant limitation to the interpretation possible in this study, as we only looked for association with an overall aggression history rather than specific types of aggression or triggers. Generalizations as to the association of these assessments to aggression in general can only be made. The internal validity of the specific subtests needs to be more closely evaluated by comparing each of them to the previously validated CBARQ factors. Additionally, the stress of being placed in a shelter might alter the dog's behavior and affect assessment results. Further investigation of test–retest reliability using dogs tested in both a home environment and a shelter might allow better measurement of the effect of stress on the results. Lastly, inter-rater reliability and intra-rater reliability are also important concepts that are under further investigation to add depth to our understanding of their interpretation.

5. Conclusions

Behavior assessments are tools commonly used in shelters to gather information about incoming dogs. Caution should be used when implementing behavior assessments that have not been thoroughly investigated as misinformation may be used to make permanent decisions concerning a dog's future. When assessments are used, the results should be used in conjunction with other types of information, such as histories obtained from relinquishing owners, and staff and volunteer observations of the animal during its stay in the facility, to create a global picture of the dog's behavior. All of this information can then be used to make informed disposition decisions for that pet and to counsel prospective adopters regarding reasonable expectations of the dog's behavior. The information can also be used to strategize the management of potential problems that may arise in the new adoptive home, thereby strengthening the developing human–animal bond.

Acknowledgements

This study was supported, in part, by a grant from the Maddie's Fund®. The Purdue Maddie's Shelter Medicine Program is underwritten by a grant from Maddie's Fund®, The Pet Rescue Foundation (www.maddiesfund.org), helping to fund the creation of a no-kill nation.

We would like to thank Lynn Tanattanawin, Lisa Lentz, and Susan Hayes for their assistance in performing and scoring the behavior assessments and in data entry.

References

American Pet Products Association, 2010. Executive Summary: APPA National Pet Owners Survey 2011–2012. American Pet Products Association, Greenwich, CT, p. 16.

Bollen, K.S., Horowitz, J., 2008. Behavioral evaluation and demographic information in the assessment of aggressiveness in shelter dogs. Appl. Anim. Behav. Sci. 112, 120–135.

Christensen, E., Scarlett, J., Campagna, M., Houpt, K.A., 2007. Aggressive behavior in adopted dogs that passed a temperament test. Appl. Anim. Behav. Sci. 106, 85–95.

Diederich, C., Giffroy, J.M., 2006. Behavioural testing in dogs: a review of methodology in search for standardization. Appl. Anim. Behav. Sci. 97, 51–72.

Feinstein, A.R., 2002. Principles of Medical Statistics. Chapman and Hall/CRC, Boca Raton, pp. p.438–451.

Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. J. Am. Vet. Med. Assoc. 223, 1293–1300.

Jones, A.C., Gosling, S.D., 2005. Temperament and personality in dogs (*Canis familiaris*): a review and evaluation of past research. Appl. Anim. Behav. Sci. 95, 1–53.

Mornement, K.M., Coleman, G.C., Toukhsati, S., Bennett, P.C., 2010. A review of behavioral assessment protocols used by Australian animal shelters to determine the adoption suitability of dogs. J. Appl. Anim. Welf. Sci. 13, 314–329.

Netto, W.J., Planta, D.J.U., 1997. Behavioural testing for aggression in the domestic dog. Appl. Anim. Behav. Sci. 52, 243–263.

Patronek, G.J., Glickman, L.T., Beck, A.M., McCabe, G.P., Ecker, C., 1996. Risk factors for relinquishment of dogs to an animal shelter. J. Am. Vet. Med. Assoc. 209, 572–581.

Salman, M.D., New, J.G., Scarlett, J.M., Kass, P.H., Ruch-Gallie, R., Hetts, S., 1998. Human and animal factors related to the relinquishment of dogs and cats in 12 selected animal shelters in the United States. J. Appl. Anim. Welf. Sci. 1, 207–226.

Segurson, S.A., 2009. Managing and rehoming the rescue dog and cat. In: Horowitz, D., Mills, D.S. (Eds.), BSAVA Manual of Canine and Feline Behavioural Medicine. , second ed. British Small Animal Veterinary Association, Quedgeley, Gloucester, pp. 270–280.

Segurson, S.A., Serpell, J.A., Hart, B.J., 2005. Evaluation of a behavioral assessment questionnaire for use in the characterization of behavioral problems of dogs relinquished to animal shelters. J. Am. Vet. Med. Assoc. 227, 1755–1761.

Steel, R.G.D., Torrie, J.H., 1960. Table A.1 Principles and Procedures of Statistics: With Special Reference to the Biological Sciences. McGraw-Hill Book Company, New York/Toronto/London, pp. 428–429.

Taylor, D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. J. Vet. Behav. 1, 94–108.

van der Borg, J.A.M., Beerda, B., Ooms, M., Silveira de Souza, A., van Hagen, M., Kemp, B., 2010. Evaluation of behaviour testing for human directed aggression in dogs. Appl. Anim. Behav. Sci. 128, 78–90.

van der Borg, J.A.M., Netto, W.J., Planta, J.U., 1991. Behavioural testing of dogs in animal shelters to predict problem behaviour. Appl. Anim. Behav. Sci. 32, 237–251.

Weiss, E., 2007. Meet Your Match SAFER^{TM} manual and training guide. ASPCA. meetyourmatch@aspca.org.